Université de Sherbrooke

ÉTUDE DES G-QUADRUPLEXES COMME RÉGULATEURS DE L'ARN

Par

Jean-Denis Beaudoin

Département de Biochimie

Thèse présentée à la Faculté de médecine et des sciences de la santé

en vue de l'obtention du grade de *philosophiae doctor* (Ph.D.) en Biochimie

Sherbrooke, Québec, Canada

Février 2013

Your file Votre référence
ISBN: 978-0-494-96340-1

Our file Notre référence
ISBN: 978-0-494-96340-1

RÉSUMÉ

ÉTUDE DES G-QUADRUPLEXES COMME RÉGULATEURS DE L'ARN

Par

Jean-Denis Beaudoin
Département de biochimie
Université de Sherbrooke

Thèse présentée à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de *philosophiae doctor* (Ph.D.) en biochimie

Avec la récente découverte que plus de 90% du génome humain est transcrit activement, il est raisonnable d'assumer que les mécanismes de régulation post-transcriptionnelle sont les moyens primaires contrôlant le transfert de l'information de l'ARN messager à la protéine. Ces mécanismes de régulation nécessitent généralement plusieurs éléments et motifs d'ARN en *cis* retrouvés à l'intérieur des ARN messagers. La structure G-quadruplexe sort de l'ordinaire en terme de motif d'ARN. L'empilement des G-quartets, formés de quatre guanines coplanaires interagissant entre elles via des paires de bases Hogsteen, la présence d'un contre-ion et la structure en tétrahélice procurent à la structure G-quadruplexe une stabilité remarquable. Cette stabilité amalgamée à ces caractéristiques structurales uniques, font de ce motif un élément de régulation post-transcriptionnelle en *cis* très prometteur. Cette thèse présente une étude des capacités de la structure G-quadruplexe à agir comme un élément de régulation de l'ARN.

Tout d'abord, j'ai exploré l'habilité d'une structure G-quadruplexe à moduler l'activité catalytique d'un ribozyme en développant et caractérisant une nouvelle classe de ribozyme, le G-quartzyme. Le G-quartzyme résulte de la fusion d'un motif G-quadruplexe au ribozyme VHD antigénomique. Une activité catalytique dépendante de la présence de potassium en solution a été observée pour ce nouveau ribozyme. La caractérisation de cette chimère G-quadruplexe-ribozyme a permis d'apprécier la flexibilité et la capacité du G-quadruplexe à moduler l'activité catalytique d'un ribozyme.

Par la suite, j'ai étudié les G-quadruplexes présents dans les 5'-UTR des ARNm en utilisant une approche robuste composée de trois étapes, *in silico*, *in vitro* et *in cellulo*. Cette méthodologie a permis d'avoir une vue d'ensemble du phénomène. L'analyse de neuf candidats de front a été la clé afin d'apprécier l'ampleur des G-quadruplexes dans les 5'-UTR agissant comme répresseurs traductionnels. Les résultats obtenus ont permis d'identifier des nouvelles règles régissant la formation de structure G-quadruplexe d'ARN *in vitro* et *in cellulo*. Ce travail suggère que ces répresseurs de la traduction sont vastement distribués à travers le transcriptome.

Finalement, cette même approche a été utilisée afin d'explorer les G-quadruplexes présents dans les 3'-UTR des ARNm. Cette analyse m'a permis de discerner un nouveau rôle pour cette structure, celui de stimuler la polyadénylation alternative d'un messager. L'étude plus en détail d'un candidat, *FXR1*, démontre que la présence d'un G-quadruplexe dans son 3'-UTR augmente l'expression d'un transcrit plus court, produit par polyadénylation alternative, contenant moins de sites de liaison aux microARNs résultant en un gain de synthèse protéique. Les résultats recueillis lors de ce travail suggèrent également que la présence de ce motif dans les 3'-UTR diminue l'efficacité d'un site de polyadénylation situé en aval de celui-ci. Clairement, les G-quadruplexes présents dans les 3'-UTR possèdent différents rôles pouvant affecter l'expression d'un gène.

En conclusion, ces études ont permis de soulever l'importance majeure des G-quadruplexes d'ARN dans différents phénomènes, dont l'expression génique, et de définir de nouvelles règles majorant leur formation et leur interaction dans divers contextes cellulaires. Les résultats présentés dans cette thèse démontrent que la structure G-quadruplexe, en plus d'être largement distribuée à travers le transcriptome, possède plusieurs caractéristiques faisant de celle-ci un élément de régulation de l'ARN des plus compétent. L'identification et la caractérisation de phénomènes cellulaires associés aux G-quadruplexes s'avèrent indispensables afin de développer de nouvelles thérapies géniques ciblant ces structures.

Mots-clés: G-quadruplexe, régulation post-transcriptonnelle, structure d'ARN, traduction, polyadénylation, expression génique.

Membres du jury d'évaluation:

Dr Jean-Pierre Perreault (Biochimie)

Dr Hervé Moine (Université de Strasbourg)

Dr Darel Hunting (Médecine nucléaire et radiologie)

Dr François Bachand (Biochimie)

Dédicace:

Je dédie cette thèse à mon oncle Dany Cantin qui a rejoint la route des étoiles en 2004. C'est chez Dany que j'ai lu mon premier article scientifique, qui traînait dans le salon, à l'âge d'environ 14 ans. Ce fût l'éclosion de mon esprit scientifique et le début de mon intérêt pour la science. Dany a été un modèle extraordinaire de courage et de persévérance. Encore aujourd'hui, je peux le ressentir et m'en nourrir. Pour ces raisons mon oncle, je te dédie ce travail.

TABLE DES MATIÈRES

LISTE DES FIGURES

Introduction

Chapitre 1

Chapitre 2

Chapitre 3

Discussions

LISTE DES TABLEAUX

Chapitre 2

Chapitre 3

LISTE DES ABRÉVIATIONS

2'-OH	2'-hydroxyle
3'	extrémité 3'
3'-UTR	*3'-UnsTraslated Region* (région 3' non traduite)
5'	extrémité 5'
5'-GMP	acide guanylique
5'-UTR	*5'-UnsTraslated Region* (région 5' non traduite)
A	adénine
ADN	acide désoxyribonucléique
anti-Akir2	effecteur oligonucléotidique négatif du G-quadruplexe *akir2*
anti-G4	effecteur oligonucléotidique négatif du ArtG4
anti-H2	effecteur oligonucléotidique négatif du G-quadruplexe *H2AFY*
APP	*Amyloid Precursor Protein* (protéine précurseur de l'amyloïde)
ARE	*AU-rich element* (élément de type AU-riche)
ARN	acide ribonucléique
ARNm	ARN messager
ARNt	ARN de transfert
ArtG4	G-quadruplexe artificiel à longue boucle 2
Ba^{2+}	ion baryum
BL	bloqueur
BS	biosenseur
C	cytosine
c	lien glycosidique
c.-d-à.	c'est-à-dire
Ca^{2+}	ion calcium
cC	*consecutive C score*
cG	*consecutive G score*
cG/cC	*consecutive G score / consecutive C score*
Cs^{+}	ion césium
CSB II	*Conserved Sequence Block II*
DC	Dichroïsme Circulaire
DMS	diméthylsulfate

ENCODE	*ENCyclopedia Of DNA elements*
FGF-2	*Fibroblast growth factor 2* (facteur de croissance des fibroblastes 2)
FMRP	*Fragile X Mental Retardation Protein*
FXR1P	*Fragile X Related Protein 1*
G	guanine
G4P	G-quadruplexe potentiel
GFP	*Green Fluorescent Protein*
GQRz	G-quartzyme
hnRNP H/F	*heterogeneous nuclear RiboNucleoProtein H/F*
IRE	*Iron Response Element* (élément de réponse au fer)
IRES	*Internal Ribosome Entry Site* (site d'entrée interne pour les ribosomes)
IRP1	*Iron-Responsive element-binding Protein 1* (protéine régulatrice du fer 1)
K^+	ion potassium
KCl	chlorure de potassium
Li^+	ion lithium
LiCl	chlorure de lithium
Mg^{2+}	ion magnésium
$MgCl_2$	chlorure de magnésium
miARN	microARN
mM	millimolaire
N	Tous les nucléotides: A, C, G ou T (ADN) ou U (ARN)
Na^+	ion sodium
NaCl	chlorure de sodium
NCBI	*National Center for Biotechnology Information*
NH_4^+	ion ammonium
NHE	*Nuclease-Hypersensitive Element*
nm	nanomètre
nM	nanomolaire
nt	nucléotide
p. ex.	par exemple
PAGE	*PolyAcrylamide Gel Electrophoresis*

PCBP	*Poly(rC)-Binding Protein*
PDB	*Protein Data Bank*
poly-A	poly-adénosines
pré-ARNm	ARN pré-messager
PUM1	Pumilio-1
Rb^+	ion rubidium
RBS	*Ribosome-Binding Site* (site d'entrée des ribosomes)
RISC	*RNA-Induced Silencing Complex*
RMN	Résonance Magnétique Nucléaire
RNase	ribonucléase
Rz	ribozyme
SAFA	*Semi-Automated Footprinting Analysis*
SELEX	*Systematic Evolution of Ligands by Exponential Enrichement*
SNP	*Single-Nucléotide Polymorphism* (polymorphisme d'un seul nucléotide)
SOFA	*Specific On/Off Adapter*
Sr^{2+}	ion strontium
ST	stabilisateur
T	thymine
T_m	*melting temperature* (température de dénaturation)
$TMPyP_4$	*meso*-5,10,15,20-tetrakis-(N-méthyle-4-pyridyle)porphine
TNFα	*Tumor necrosis factor-alpha* (facteur de nécrose tumorale alpha)
TRAP	*Telomeric Repeat Amplification Protocol*
U	uracile
VDH	virus de l'hépatite D
VHC	virus de l'hépatite C
wt	*wild type* (type sauvage)
µM	micromolaire

INTRODUCTION

1. Structures des acides nucléiques

L'ADN (acide désoxyribonucléique), macromolécule connue de tous et symbole même de la vie, serait présente sur terre depuis plusieurs milliards d'années. Toutefois, il faut remonter à l'hiver 1868-69 avant que l'espèce humaine ne prenne connaissance de son existence. Un étudiant du laboratoire de Felix Hoppe-Seyler de l'Université de Tübingen, Friedrich Miescher, réussit à isoler des noyaux de leucocytes une substance qui était ni protéique ni lipidique et qui ne ressemblait à rien de connu à cette époque (Dahm, 2008). Elle était composée de carbone, d'azote et de phosphore. Il décida de la nommer "nucléine". C'est seulement en 1889 que le nom "acide nucléique" fût attribué à la "nucléine" par Richard Altmann due à son caractère acide. Dans ce même laboratoire, un homme dénommé Albrecht Kossel poursuivit les recherches sur cette "nucléine" et entreprit de déterminer sa structure. Il découvrit les cinq bases, soient: l'adénine, la cytosine, la guanine, la thymine et l'uracile. Une autre percée importante eut lieu en 1919 alors que l'unité de base de l'acide nucléique fût découverte. Phoebus Aaron Theodor Levene identifia le nucléotide et ses trois composantes: la base, le sucre et le groupement phosphate. Il découvrit également que ces nucléotides étaient reliés par un lien covalent entre le groupement phosphate et le sucre formant en réalité un polymère (Tipson, 1957). Plus important encore dans le cadre de cette thèse, il identifia les différents sucres pouvant composer l'acide nucléique, le ribose et le désoxyribose, permettant de distinguer pour la première fois l'ARN (acide ribonucléique) de l'ADN.

1.1 La double hélice d'ADN

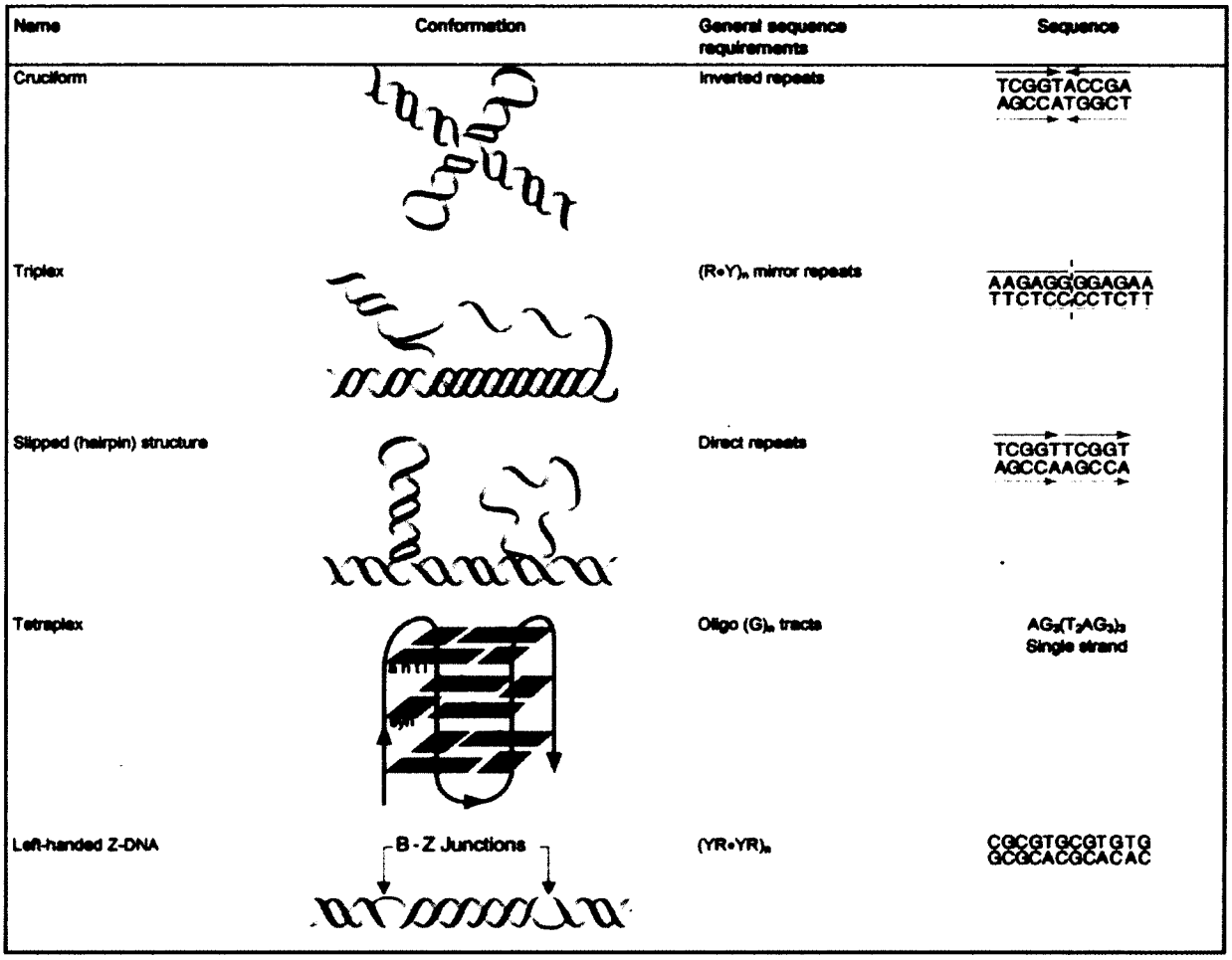
Mainte fois qualifié de la plus grande découverte du 20e siècle, la résolution de la structure double hélice de l'ADN a eu l'effet d'une bombe dans le monde scientifique. En 1953, les travaux de Francis Crick, James Watson, Maurice Wilkins et Rosalind Franklin ont permis de percer le mystère de la structure de

l'ADN et la double hélice d'ADN fût révélée au grand jour (Watson and Crick, 1953). Cette découverte, représentant rien de moins que l'origine de la biologie moléculaire comme nous la connaissons aujourd'hui, a valu à ses découvreurs le prix Nobel de physiologie ou de médecine en 1963 (à l'exception de Franklin, décédée quelques années plus tôt). Lors d'une présentation historique en 1957, Crick décrit ce qui est devenu le célèbre "dogme central" de la biologie moléculaire (Crick, 1970). Les mécanismes généraux de ce "dogme" comportent la réplication de l'ADN ainsi que le code génétique qui est transcrit en ARN, puis traduit en protéines par l'utilisation de codons correspondants à des triplets de nucléotides. La structure double hélice et sa complémentarité de bases (A-T et G-C) font de celle-ci la macromolécule de choix pour contenir l'information du génome cellulaire et assurer une transmission fidèle.

1.2 Des structures inhabituelles et inusitées

Alors que la double hélice d'ADN (ADN de pas droit et de type B) prit pratiquement toute la place sous les projecteurs, c'est dans l'ombre de celle-ci que d'autres types de structures d'ADN (de type non B) émergèrent. Bien que plusieurs de celles-ci n'ont été observées uniquement *in vitro*, d'autres, au contraire, peuvent également se former *in vivo* et même affecter l'homéostasie cellulaire (Wells, 2007). À titre d'exemples de structures d'ADN de type non B retrouvées *in vivo*, il y a le cruciforme (formé par des séquences répétées inversées), la triple hélice ou triplex (formée par des répétitions miroirs), la structure tige boucle causée par le glissement de la double hélice (formée par des répétitions directes), la double hélice de pas gauche de type Z (formée par des séries de doublets purine-pyrimidine) et, finalement, le G-quadruplexe (formé par plusieurs séries de guanines consécutives) (Intro, Figure 1). Ces structures diffèrent de la double hélice d'ADN de type B par des angles de liaison contorsionnés ou des régions non appariées. La machinerie cellulaire n'étant pas généralement adéquatement équipée pour gérer ces répétitions de structures inusitées, ces dernières entraînent souvent l'apparition d'erreurs dans le génome menant parfois au développement

de maladies connues (la dystrophie myotonique, la maladie d'Huntington, le syndrome du X fragile et l'ataxie de Friedreich pour nommer quelques exemples)(Mirkin, 2007; Bacolla and Wells, 2009).



Intro, Figure 1. Conformations d'ADN de type non B.

Différentes conformations d'ADN de type non B provoquant souvent de l'instabilité génétique et des réarrangements chromosomiques. Un schéma est représenté pour chacune des conformations tout comme les conditions typiquement requises au niveau de la séquence primaire pour mener à leur formation. Un des brins du duplex d'ADN est représenté en bleu alors que le brin complémentaire est représenté en rouge. Figure tirée de (Wells, 2007).

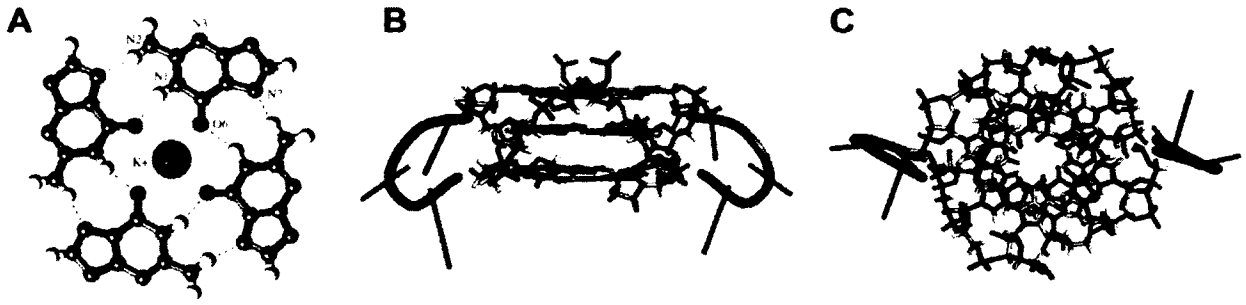
2. La structure G-quadruplexe

En 1910, plus d'une quarantaine d'années avant que Watson et Crick proposent la double hélice d'ADN, le chimiste allemand Ivar Bang observa que l'acide guanylique (5'-GMP) formait des sortes de gels à des concentrations de l'ordre du haut millimolaire. Cette propriété physique étrange intrigua les scientifiques jusqu'en 1962, où Gellert et ses associés percèrent le secret de ce phénomène (Gellert et al., 1962). En analysant des spectres de diffraction aux rayons X de fibres formant ces gels, ils découvrirent qu'une unité tétramérique bien particulière s'assemblait les unes avec les autres façonnant une grande structure hélicoïdale. Cette unité tétramérique est maintenant appelée G-quartet ou G-tétrade, et sera décrite ci-dessous, tandis que la grande structure tétrahélicoïdale formée par l'empilement de G-quartets correspond à la structure G-quadruplexe. Les G-quadruplexes demeureront pendant plus d'une vingtaine d'années des curiosités *in vitro* intéressant majoritairement les biophysiciens. Ceux-ci détermineront un fait marquant pour cette thèse, c'est-à-dire qu'autant l'ADN que l'ARN peuvent former une structure G-quadruplexe (Chantot et al., 1971). Cependant, vers la fin des années 1980, l'univers des G-quadruplexes fût complètement chamboulé par une série de publications conférant à cette structure d'ADN des rôles biologiques potentiels. Les premières séquences d'ADN riches en guanines à connotation biologique à avoir été démontrées pour former des G-quadruplexes stables furent les séquences télomériques (Henderson et al., 1987; Sundquist and Klug, 1989) et la portion riche en guanines retrouvée dans la région variable des gènes de l'immunoglobuline (Sen and Gilbert, 1988). Depuis cette étape marquante, une multitude de structures G-quadruplexes ont été identifiées pour être impliquées dans plusieurs mécanismes et processus cellulaires.

2.1 Le G-quartet

Le G-quartet est l'unité de base du G-quadruplexe et il est constitué de quatre guanines coplanaires qui interagissent les unes avec les autres par l'intermédiaire de paires de bases Hoogsteen. Comme chacune des guanines oriente son

groupement carbonyle O_6, chargé partiellement négativement, au centre du G-quartet, sa formation dépend de la présence d'un contre-ion, généralement un cation monovalent (Intro, Figure 2A)(Gellert et al., 1962). Par la suite, plusieurs G-quartets s'empilent les uns au dessus des autres afin de former une structure tétrahélicoïdale, le G-quadruplexe (Intro, Figure 2B et C).



Intro, Figure 2. G-quartet et G-quadruplexe.

(**A**) Représentation d'un G-quartet mettant en évidence le réseau de ponts hydrogènes formé entre les faces Hoogsteen et Watson-Crick des différentes guanines. Le cation potassium est retrouvé au centre du G-quartet (sphère mauve). Les sucres des guanosines ont été omis pour une meilleure clarté. (**B**) et (**C**) Vue de côté ou de dessus, respectivement, du G-quadruplexe formé par la séquence télomérique d'ARN bimoléculaire en présence de potassium. L'analyse structurale a été réalisée avec le programme PyMOL en utilisant la structure 2KBP de la base de données "Protein Data Bank" (PDB). Le code de couleur des atomes est le suivant: le carbone (gris), l'oxygène (rouge), l'azote (bleu), l'hydrogène (blanc) et le phosphore (orange). Figure adaptée de (Neidle and Balasubramanian, 2006).

Deux propriétés particulières font de la guanosine (Intro, Figure 3A) le nucléotide de prédilection dans le but de construire un tel échafaudage. Premièrement, elle possède une distribution unique de groupements donneurs et accepteurs de ponts hydrogènes autant sur son côté Watson-Crick que Hoogsteen (Intro, Figure 3B). Chacune des guanosines du G-quartet accepte deux ponts hydrogènes des groupements NH (position N_1) et NH_2 (position N_2) d'une des guanosines adjacentes via ses groupement O_6 et N_7, respectivement, et agit comme donneur de deux ponts hydrogènes via ses groupements NH (position N_1) et NH_2 (position N_2) pour les groupements O_6 et N_7, respectivement, de la seconde

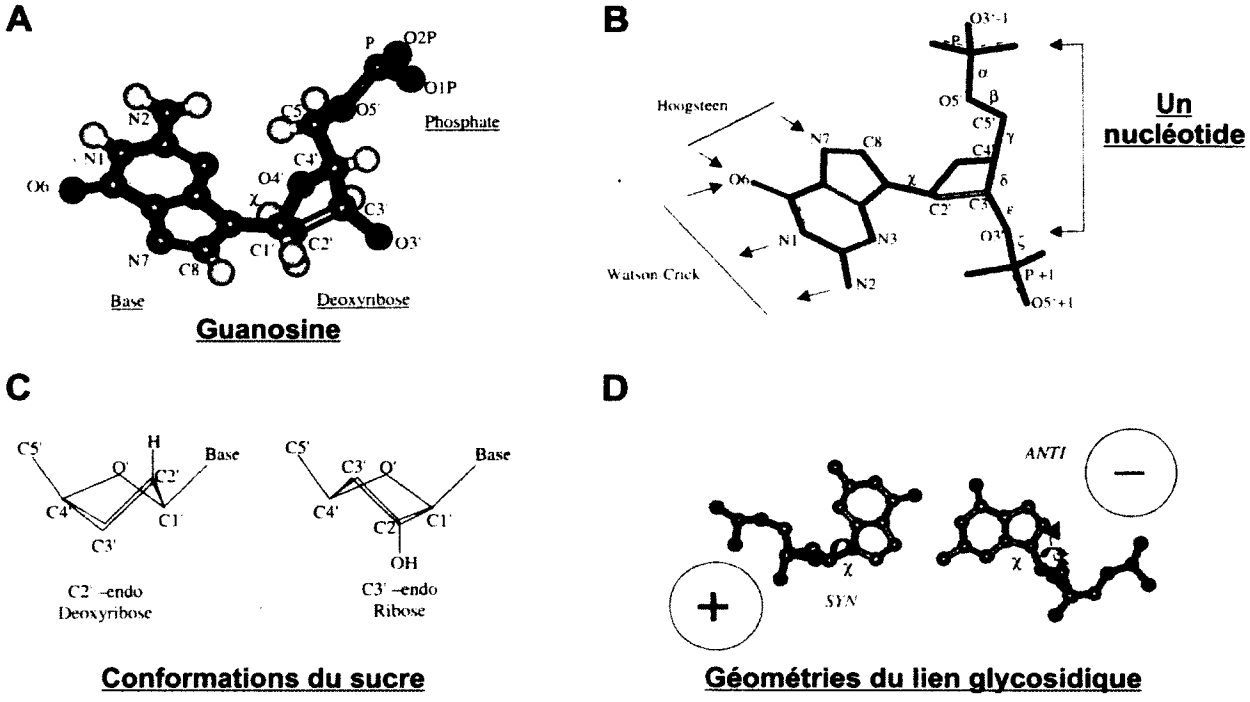
guanosine adjacente (Intro, Figure 2A). Cette organisation mène à la formation de huit ponts hydrogènes par G-quartet. Deuxièmement, son cycle aromatique possède un fort dipôle favorisant l'empilement de type π-π des G-quartets menant à la formation du G-quadruplexe (Intro, Figure 2B et C).

Outre la section base du nucléotide, la nature du sucre de ce dernier a également un impact important sur la structure du G-quartet et du G-quadruplexe. Le pentose possède une structure non planaire avec différentes conformations énergétiquement favorables. Dans les duplex et quadruplex d'ADN et d'ARN ce sont les conformations $C_{2'}$-*endo* et $C_{3'}$-*endo* qui sont les plus souvent retrouvées (Intro, Figure 3C). Le type de conformation aura un impact direct sur le positionnement du lien glycosidique (reliant le sucre à la base; Intro, Figure 3D). Il existe différentes géométries, caractérisées par l'angle de torsion du lien glycosidique (χ), adoptées par le pentose et la base. Les deux conformations les plus fréquentes concernant l'ADN et l'ARN sont *syn* (avec un angle $0 < \chi < 90°$) et *anti* (avec un angle $-120 > \chi > 180°$) (Intro, Figure 3D).

2.2 Le contre-ion positif

Comme mentionné précédement, la formation d'un G-quartet entraine le positionnement du groupement O_6 de chacune des guanines vers le centre de la tétrade (Intro, Figure 2A). Cette proximité de quatre groupements chargés partiellement négativement devrait être très défavorable à la formation du G-quartet. Cette réalité fait place à l'une des caractéristiques les plus importantes des G-quadruplexes, c'est-à-dire la nécessité d'avoir un contre-ion positif capable de coordonner ces charges négatives (Arnott et al., 1974). Ce n'est pas n'importe lequel ion qui peut jouer ce rôle stabilisateur et la structure G-quadruplexe a ses propres préférences. Le contre-ion doit avoir une charge positive adéquate, une taille particulière et une énergie de déshydratation la plus faible possible. L'ion potassium (K^+) est le contre-ion idéal et celui le plus souvent retrouvé à l'intérieur des G-quadruplexes, suivi dans l'ordre des ions ammonium (NH_4^+), rubidium (Rb^+),

sodium (Na^+) et césium (Cs^+) (Wong and Wu, 2003). Dans certaines conditions, des ions divalents comme le strontium (Sr^{2+}), le baryum (Ba^{2+}), le calcium (Ca^{2+}) et le magnésium (Mg^{2+}) peuvent également aider à la formation de G-quadruplexes (Venczel and Sen, 1993). À l'inverse, l'ion lithium (Li^+) est parfaitement incapable de promouvoir la formation de G-quadruplexes, majoritairement à cause de sa trop petite taille.

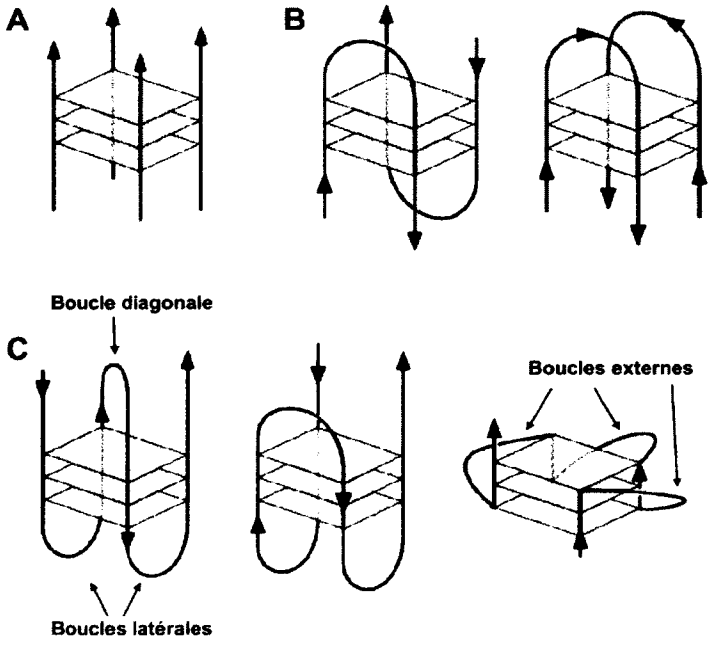


Intro, Figure 3. Les guanosines du G-quadruplexe.

(**A**) Schéma atomique d'une guanosine avec ses trois composantes: un groupement phosphate, un sucre (ici un désoxyribose) et une base (la guanine). (**B**) Deuxième représentation de la guanosine illustrant les faces Hoogsteen et Watson-Crick de la guanine. Les flèches désignent les groupements donneurs ou accepteurs de pont hydrogène (**C**) Schéma de la conformation $C_{2'}$-*endo* du désoxyribose et $C_{3'}$-*endo* du ribose. (**D**) Représentation de guanines mettant l'emphase sur l'angle de torsion du lien glycosidique reliant le sucre et la guanine. Le code de couleur des atomes: le carbone (gris), l'oxygène (rouge), l'azote (bleu), l'hydrogène (blanc) et le phosphore (orange). Figure adaptée de (Neidle and Balasubramanian, 2006).

2.3 L'hétérogénéité de topologies

En présence d'un contre-ion approprié, la structure G-quadruplexe peut se former à partir d'un ou de plusieurs oligonucléotides d'ADN ou d'ARN riche en guanosines (Intro, Figure 4). Dans le cadre de cette thèse, l'emphase sera portée sur les G-quadruplexes unimoléculaires formés à partir d'une seule molécule d'ADN ou d'ARN. Les différentes topologies formées par les G-quadruplexes peuvent se séparer en deux grandes catégories: parallèles et antiparallèles (Burge et al., 2006). Les G-quadruplexes parallèles ont la particularité que chaque brin de leur tétrahélice est orienté dans la même direction (soit 5'-3' ou 3'-5'). À l'inverse, les G-quadruplexes antiparallèles ont au moins un de leur brin qui est orienté dans une direction opposée de celle des autres (Intro, Figure 4). Ces orientations différentes entrainent un réarrangement géométrique obligeant certaines guanosines à passer d'une conformation *anti* à *syn* dans le but de maintenir l'agencement des G-quartets. Les G-quadruplexes antiparallèles sont donc caractérisés par un mélange de guanosines de conformation *anti* et *syn*, tandis que les parallèles sont homogènes pour la conformation *anti*. Cette différence permet de les discriminer avec la technique de dichroïsme circulaire, entre autres, comme on le verra plus loin.



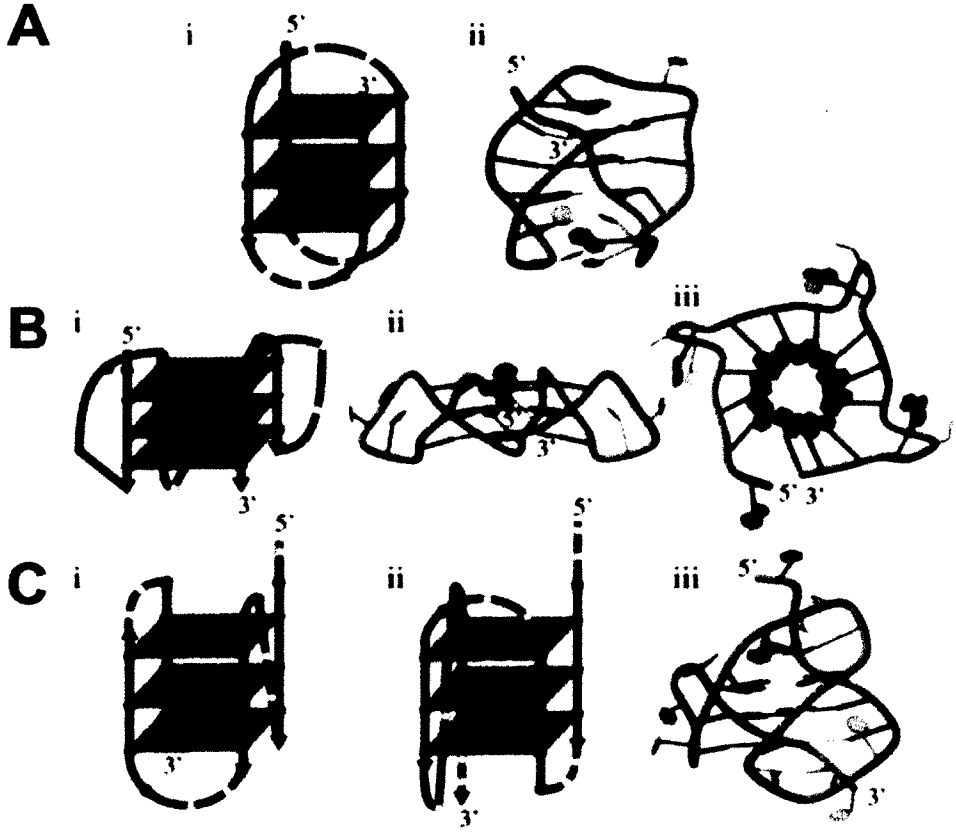
Intro, Figure 4. Topologies des G-quadruplexes.

(**A**) G-quadruplexe parallèle tétramoléculaire. (**B**) Deux G-quadruplexes antiparallèles bimoléculaires possédant deux boucles latérales organisées de façon tête-à-queue (à gauche) ou tête-à-tête (à droite). (**C**) Trois G-quadruplexes unimoléculaires dont deux de types antiparallèles (à gauche et au centre) et un de type parallèle (à droite). Le G-quadruplexe de gauche possède deux boucles latérales et une boucle diagonale, alors que celui du centre comprend trois boucles latérales. Le G-quadruplexe parallèle (à droite) possède quant à lui trois boucles externes. Figure adaptée de (Neidle and Balasubramanian, 2006).

L'orientation des différents brins aura un impact majeur sur le type de boucle observé pour un G-quadruplexe donné. Les boucles correspondent aux séquences nucléotidiques reliant les différents brins. Il y a trois types principaux de boucles: la diagonale (reliant une guanosine d'un G-quartet terminal à la guanosine opposé du même G-quartet, Intro, Figure 4C), la latérale (reliant une guanosine d'un G-quartet terminal à une des guanosines adjacentes du même G-quartet; Intro, Figure 4C) et l'externe ou "*propeller*" (reliant une guanosine d'un G-quartet terminal à une guanosine de l'autre G-quartet terminal; Intro, Figure 4C). Dans le cadre d'un G-quadruplexe unimoléculaire, uniquement la présence de trois boucles externes peuvent mener à la formation d'une topologie parallèle puisque les boucles diagonales et latérales entrainent nécessairement un changement d'orientation des brins à l'intérieur de la structure.

Il est très important de noter qu'en raison de leur ribose de conformation $C_{3'}$-*endo*, forçant un lien glycosidique de géométrie *anti*, les G-quadruplexes d'ARN sont limités à la formation de structures parallèles composées de trois boucles externes. Toutefois, les G-quadruplexes à base d'ADN ont une plus grande diversité de topologies. Le choix de topologie adoptée est influencé par la séquence d'acide nucléique et les conditions expérimentales utilisées comme la nature du contre-ion et la concentration d'acide nucléique. La séquence d'ADN télomérique comprenant quatre répétitions de série de guanosines est un exemple parfait de cette diversité structurale (Bryan and Baumann, 2011). Pas moins de cinq structures de conformations différentes ont été résolues par résonance

magnétique nucléaire (RMN) et l'étude de cristaux par diffraction de rayons X (dont quatre sont représentés en Intro, Figure 5). De plus, il est fréquent d'observer un mélange de topologies pour une séquence donnée dans une même condition.



Intro, Figure 5. G-quadruplexes télomériques intramoléculaires.

Les guanosines, les adénosines et les thymidines sont représentées par des sphères de couleur mauve, verte et jaune respectivement. (**A**) Topologie (*i*) et structure RMN (*ii*) de l'oligonucléotide $AGGG(TTAGGG)_3$ dans une solution de sodium formant un G-quadruplexe antiparallèle avec deux boucles latérales et une diagonale. (**B**) Topologie (*i*) et la structure cristalline (*ii*, vue de cotée et *iii*, vue de dessus) de l'oligonucléotide $AGGG(TTAGGG)_3$ dans une solution contenant du potassium formant un G-quadruplexe parallèle avec trois boucles externes. (**C**) Les topologies hybrides retrouvées en solution contenant du potassium. Les topologies hybride 1 (*i*) et hybride 2 (*ii*) illustrant différentes organisations de boucles latérales et externes. La structure RMN de l'hybride 2 est montrée en (*iii*). Figure tirée de (Bryan and Baumann, 2011).

2.4 Une stabilité impressionnante

Une autre caractéristique étonnante de la structure G-quadruplexe est sa stabilité, qui peut atteindre des niveaux phénoménaux. Par exemple, à des concentrations physiologiques de cations, certains G-quadruplexes possèdent un temps de demi-vie calculé en jours à une température aussi élevée que 60°C, alors que d'autres sont toujours partiellement formés à des températures de 95-100°C (Lu et al., 1992; Gupta et al., 1993). Cette stabilité découle de chacune de ses composantes. i) L'assemblage de chaque G-quartet implique la formation d'un grand total de huit ponts hydrogènes. ii) L'empilement des G-quartets les uns par-dessus les autres, résultant en la formation d'une tétrahélice, engage une combinaison de forces hydrophobiques, électrostatiques et de van der Waals. iii) Comme toute structure hélicoïdale, le G-quadruplexe possède des sillons, au nombre de quatre pour cette tétrahélice. À l'intérieur de ceux-ci, on retrouve un réseau de molécules d'eau ordonné par la présence de plusieurs groupements donneurs et accepteurs de ponts hydrogènes. Ce réseau hydrate en quelque sorte le G-quadruplexe. iv) La présence d'un contre-ion positif coordonnant les charges négatives à l'intérieur du G-quadruplexe participe également à la stabilité globale de la structure. Toutes ces propriétés confèrent des stabilités titanesques à certaines structures G-quadruplexes.

Il est important de noter, particulièrement dans le cadre de cette thèse, que les G-quadruplexes d'ARN sont généralement plus stables que leur contrepartie en ADN. Ce gain de stabilité est en partie dû à leur limitation à adopter uniquement des structures de type parallèle. Ces dernières possèdent, à l'inverse de celles de type antiparallèle, quatre sillons symétriques permettant un réseau d'hydratation uniforme conférant une meilleure stabilité (Neidle and Balasubramanian, 2006). De plus, la présence de groupements hydroxyles, au niveau des riboses, permet la formation d'un réseau de ponts hydrogènes plus complexe et élargi. Ces

différences font des G-quadruplexes d'ARN de véritables champions en termes de stabilité.

3. Techniques pour l'étude des G-quadruplexes

Au cours des années, plusieurs méthodologies et techniques expérimentales ont été développées dans le but d'étudier la formation de G-quadruplexes, chacune examinant différents aspects de cette structure distincte. La majorité de ces techniques sont principalement descriptives alors qu'une analyse à haute résolution par RMN ou cristallographie aux rayons X est nécessaire pour obtenir une structure complète. Voici une brève description de quelques-unes de ces techniques qui sont importantes dans le cadre de cette thèse.

3.1 Cristallographie aux rayons X et RMN

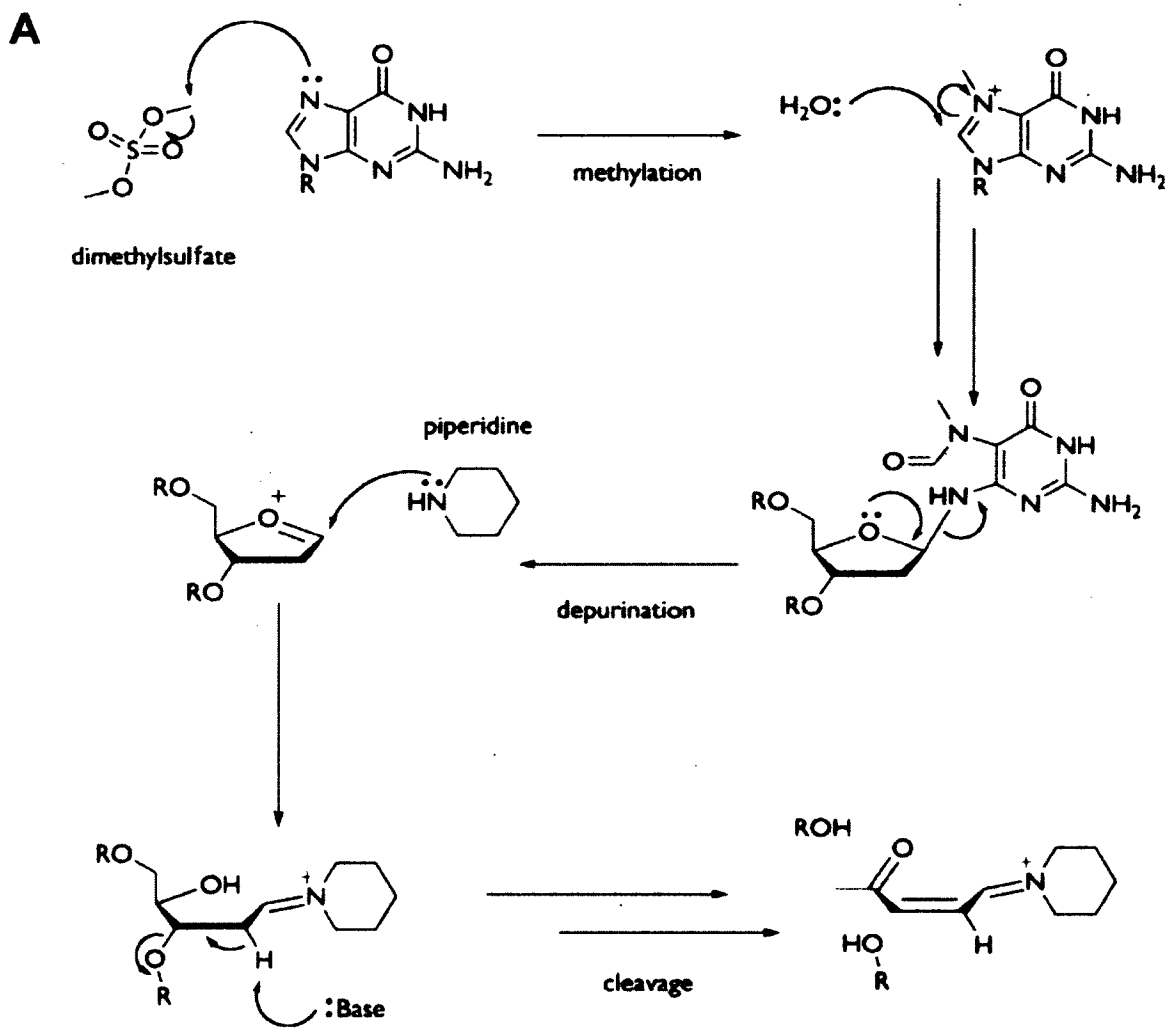
La cristallographie aux rayons X a été la première technique à fournir une structure complète d'un G-quadruplexe (Gellert et al., 1962; Zimmerman et al., 1975). À ce jour, c'est plus d'une centaine de structures G-quadruplexes à haute résolution obtenues par cristallographie ou RMN qui se retrouvent dans la base de données "*Protein Data Bank*" (PDB). Certaines structures démontrent une résolution se positionnant en dessous du 1 Å, renfermant du même coup une vraie mine d'or d'informations au niveau architectural. Si de son côté, la technique de cristallographie rapporte uniquement une structure adoptée en phase solide (Campbell and Parkinson, 2007), la technique RMN nous renseigne sur une structure formée en solution (Webba da Silva, 2007). Malgré leur grande utilité, ces deux techniques possèdent également certains défauts. Un désavantage majeur de la cristallographie est la nécessité de produire un cristal unique de la structure étudiée. Conséquemment, si une séquence forme un méli-mélo de structures (topologies) en solution, cette technique décrira uniquement celle qui se cristallise le mieux. Celle-ci peut représenter qu'une petite fraction du mélange et être bien différente de la structure retrouvée de façon majoritaire. C'est particulièrement préoccupant sachant que les G-quadruplexes sont des structures plutôt

polymorphes. De plus, ces deux techniques nécessitent généralement l'utilisation de séquences hautement mutées par la substitution de certains nucléotides de la séquence de type sauvage (p. ex. l'utilisation de 5-bromo-thymidine en cristallographie ou la substitution de la thymidine par l'uracile ou le 5-bromo-uracil dans le cas du RMN). L'emploi de ces différents analogues est nécessaire dans le but d'aider à la cristallisation d'une séquence ou d'aider à rendre plus homogène le nombre de topologies adoptées par une structure en solution. À cause des divergences expérimentales entre ces deux techniques, il est fréquent d'obtenir des structures très différentes pour une même séquence lorsque l'on compare les résultats obtenus entre celles-ci. La panoplie de structures résolues à partir de la séquence télomérique humaine représente bien cette réalité (Intro, Figure 5). C'est pourquoi l'utilisation de techniques qui étudient le mélange de structures formées dans des conditions plus "standard" est nécessaire afin de supporter et soutenir les structures à hautes résolutions obtenues. Généralement, ces dernières analyseront une caractéristique bien précise de la structure G-quadruplexe.

3.2 Cartographie à l'aide du diméthylsulfate

Cette technique utilise la propriété du diméthylsulfate (DMS) à ajouter un groupement méthyle en position N_7 de la guanine entrainant une dépurination de la guanosine. Après le traitement au DMS, l'ajout de piperidine engendre le clivage de la chaîne nucléotidique au niveau de ces différents sites abasiques (Intro, Figure 6). Les produits de clivage sont ensuite séparés par électrophorèse et le niveau de clivage au niveau de chacune des guanosines de la séquence d'intérêt est quantifié (Huppert, 2008). La formation de G-quadruplex est caractérisée par une diminution du clivage au niveau des guanosines impliquées dans la composition des différents G-quartets. En effet, le groupement N_7 des guanines des G-quartets est impliqué dans l'élaboration d'un pont hydrogène, empêchant l'ajout d'un groupement méthyle lors du traitement au DMS (Intro, Figure 6). De leur côté, les interactions Watson-Crick ne tirent pas profit de la position N_7 de la

guanine permettant la distinction des G-quadruplexes par rapport aux structures formées à partir de paires de bases Watson-Crick.

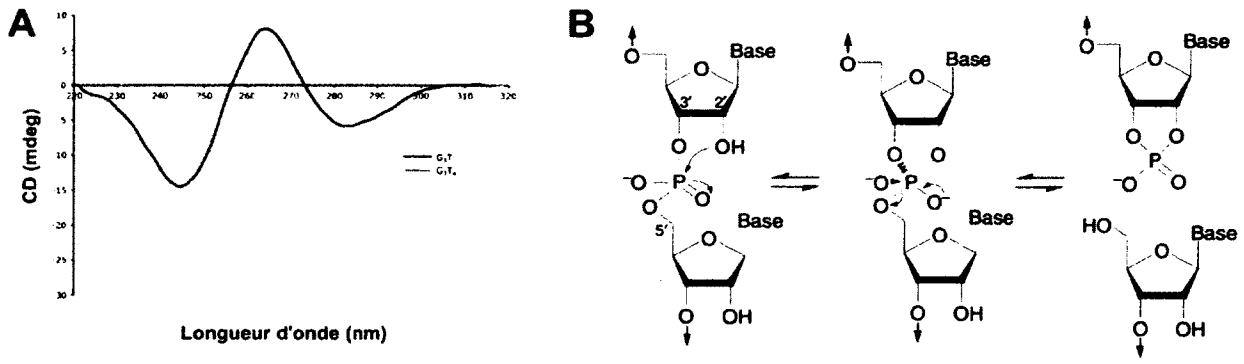


Intro, Figure 6. Cartographie au diméthylsulfate.

(**A**) Réaction engendrée suite au traitement au diméthylsulfate suivi de l'addition de piperidine menant au clivage de l'acide nucléique au niveau des guanines. La protection du groupement N_7 via la formation d'un pont hydrogène, comme c'est le cas dans la structure G-quadruplexe, empêche le clivage à l'endroit de ces guanines. Figure tirée de (Huppert, 2008).

3.3 Dichroïsme circulaire

En plus d'identifier la formation de G-quadruplexes, cette technique permet de distinguer la présence de G-quadruplexes de types parallèles et antiparallèles. Cette technique met à profit la possibilité de polariser de façon circulaire une onde lumineuse: une partie circulaire droite et une autre circulaire gauche (Solomon, 1999). La lumière polarisée est dirigée à travers une solution qui contient une molécule ou des molécules dites chirales, c'est-à-dire qu'elles interagiront de façon asymétrique avec les différentes formes énantiomériques de la lumière. Puisque ces asymétries varient en fonction de la longueur d'onde, il est possible de tracer des graphiques à partir des valeurs obtenues. La structure G-quadruplexe possède une chiralité caractéristique qui se définie par des graphiques de dichroïsme circulaire (DC) arborant certains aspects typiques (Burge et al., 2006; Huppert, 2008). Cette chiralité provient majoritairement de la tétrahélice formée et des types de liens glycosidiques (*syn* ou *anti*) présents à l'intérieur de celle-ci. Un G-quadruplexe parallèle, possédant uniquement des guanosines de conformation *anti*, sera caractérisé par la présence d'un pic positif à une longueur d'onde d'environ 264 nanomètre (nm) et d'un pic négatif à environ 240 nm. De son côté, un G-quadruplexe antiparallèle, possédant des guanosines de conformation *anti* et *syn*, sera caractérisé par la présence d'un pic positif près de 295 nm et d'un pic négatif à environ 260 nm (Intro, Figure 7A). Dans le cadre de cette thèse, l'analyse par DC porte uniquement sur des G-quadruplexes d'ARN, donc de types parallèles. Il faut prendre note qu'il existe d'autres éléments des acides nucléiques menant à la présence de graphique arborant un pic positif aux alentours de 264 nm. Il faut alors redoubler de prudence lors de l'analyse des résultats. Plutôt que de se fier à l'analyse d'un seul graphique, il est préférable de comparer les résultats obtenus dans des conditions ne permettant pas la formation de G-quadruplexe (c.-à-d. en absence de contre-ion ou en présence de Li^+) à ceux enregistrés dans des conditions favorables (c.-à-d. en présence de Na^+ ou K^+). Si une transition plus prononcée vers un graphique possédant les aspects notoires d'un G-quadruplexe parallèle est observée, on peut fortement suggérer que la séquence analysée forme bel et bien cette structure dans les conditions testées.



Intro, Figure 7. Dichroïsme circulaire et "*in-line probing*".

(**A**) Spectres de dichroïsme circulaire caractéristiques d'une structure G-quadruplexe parallèle (trait foncé) et antiparallèle (trait pâle). (**B**) Réaction produite lors d'un clivage résultant d'une attaque nucléophile du groupement 2'-hydroxyle (2'-OH). Quand le 2'-OH du ribose entre dans une conformation linéaire en ligne directe avec le groupement phosphate et l'oxygène en 5' du ribose précédant, celui-ci exécute une attaque nucléophile efficace sur le phosphate menant au clivage de la molécule d'ARN à cet endroit. Figure tirée de (Huppert, 2008) et (Regulski and Breaker, 2008).

3.4 La dénaturation thermique

La dénaturation thermique est un processus où il est possible de dénaturer certaines structures d'acide nucléique en faisant simplement varier la température. L'objectif est de commencer avec une température pour laquelle la structure étudiée est stable et de l'augmenter vers une valeur pour laquelle la structure devient instable. Lors de cette dénaturation par l'augmentation de la température, il est possible de suivre le stade de dénaturation par l'enregistrement de différentes valeurs (Mergny and Lacroix, 2009). Dans le cas des G-quadruplexes d'ARN (parallèle), il est possible de surveiller leur dénaturation en suivant leur valeur de DC aux alentours de 264 nm reflétant leur formation (pour les raisons mentionnées précédemment). En analysant les données de DC obtenues en fonction de la température, il est facile de calculer une température de dénaturation (T_m, température à laquelle la moitié des structures sont dénaturées) pour une condition donnée (Mergny and Lacroix, 2009). La formation de G-quadruplexes est caractérisée par une augmentation de la valeur de T_m dans des conditions où cette

structure se forme, généralement en présence de Na^+ et/ou de K^+. Il arrive parfois, et c'est souvent le cas pour les G-quadruplexes d'ARN, qu'aucune valeur de T_m ne puisse être calculée pour une séquence, car la structure est tellement stable qu'elle n'est pas encore complètement dénaturée à des températures aussi élevées que 95°C.

3.5 "In-line probing"

Cette technique de cartographie chimique de structure secondaire d'ARN est probablement l'une des plus simples qui existe (Regulski and Breaker, 2008). Elle exploite la tendance naturelle de l'ARN à se dégrader en fonction de sa structure. Les liens phosphodiesters qui relient de façon covalente chacun des ribonucléotides d'une molécule d'ARN sont sujets à un clivage lent et non enzymatique. Celui-ci est dicté par l'agencement en ligne directe des groupements 2'-hydroxyle (2'-OH) du ribose ainsi que du phosphate et de l'oxygène en position 5' du ribose précédent. Lorsque ces groupements adoptent cette géométrie en ligne directe, le 2'-OH agit en tant que nucléophile permettant le déplacement intramoléculaire de l'oxygène en position 5' vers le groupement phosphate adjacent menant au clivage de l'ARN à un endroit précis (Intro, Figure 7B). Le niveau de clivage dépend de: i) l'état d'ionisation du groupement 2'-OH; ii) de la distance entre le 2'-OH et le phosphate; et, iii) le bon agencement en ligne directe de tous les groupements impliqués (Soukup and Breaker, 1999). Suivant cette logique, les nucléotides flexibles et simples brins libres d'adopter un grand nombre de géométries différentes, incluant celle en ligne directe, subissent plus régulièrement ce clivage spontané. À l'inverse, les nucléotides contraints à une conformation précise à l'intérieur de la structure comme une double ou une tétrahélice sont moins propices à adopter cette géométrie en ligne directe limitant du même coup leur clivage. Afin de favoriser ce clivage spontané, les molécules d'ARN à analyser sont incubées pendant 40 heures à la température de la pièce dans une solution à un pH légèrement basique et à une concentration relativement élevée de Mg^{2+}. Les produits de clivage sont ensuite séparés par électrophorèse

permettant une analyse qualitative et quantitative des patrons de bandes obtenues. Cette technique est utilisée régulièrement afin de cartographier les changements au niveau de la structure secondaire de différents riborégulateurs en présence et en absence de ligand (Regulski and Breaker, 2008). La liaison du ligand au riborégulateur entraîne un réarrangement au niveau de la structure secondaire de ce dernier. Cette réorganisation se reflète par un changement au niveau du patron de clivage et peut donc être suivi de près avec cette technique. Au cours de mes travaux de recherches, nous avons démontré pour la première fois que cette technique peut également être utilisée de manière efficace, reproductible et informative afin d'étudier la formation de G-quadruplexes unimoléculaires d'ARN.

4. La prédiction de G-quadruplexes

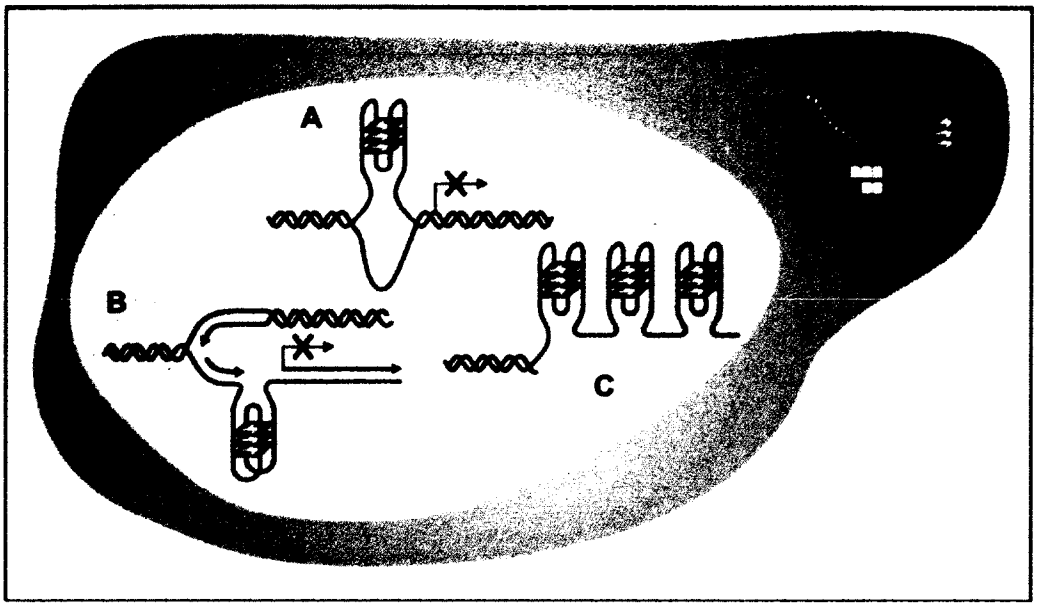
Comme il a été mentionné préalablement, à la fin des années 1980 quelques études clées suggérant certains rôles biologiques à la structure G-quadruplexe initièrent une ruée vers l'analyse des G-quadruplexes retrouvés à l'intérieur des génomes. Les scientifiques se sont d'abord portés sur l'étude de séquences extrêmement riche en guanosines ou possédant clairement plusieurs séries de guanosines consécutives (p. ex. certaines régions sujettes à des événements de recombinaison et les télomères)(Simonsson, 2001). En même temps, ils développèrent un intérêt marqué pour trouver les plus petits motifs possibles capables de former un G-quadruplexe. Une kyrielle d'études *in vitro* ont alors été réalisées à cette fin en faisant varier individuellement chacun des paramètres composant la structure G-quadruplexe dans le but d'identifier certaines règles. Au mois de mai 2005, dans le même numéro de la revue "Nucleic Acid Research", deux études rapportèrent pratiquement le même algorithme, qu'ils utilisèrent afin d'identifier plusieurs séquences possédant un fort potentiel à former une structure G-quadruplexe dans le génome humain (Huppert and Balasubramanian, 2005; Todd et al., 2005). Le concept important de "G-quadruplexe potentiel" prit vie. On réfère à un "G-quadruplexe potentiel" (G4P) une séquence identifiée uniquement

par bioinformatique comme ayant de fortes probabilités d'adopter une structure G-quadruplexe. L'algorithme utilisé parcours la séquence primaire d'ADN en question à la recherche de motifs répondant aux critères suivants: G_x-N_{1-7}-G_x-N_{1-7}-G_x-N_{1-7}-G_x, où $x \geq 3$ et N correspond à n'importe lequel des quatre nucléotides (A, C, T ou G). Avec l'aide de cette formule, un total de 376 000 G4P ont été identifiés dans le génome humain (Huppert and Balasubramanian, 2005). Aujourd'hui, plusieurs logiciels sont disponibles gratuitement sur le web afin d'utiliser ce type d'algorithme sur une séquence d'ADN ou d'ARN d'intérêt (Kikin et al., 2006; Scaria et al., 2006; Menendez et al., 2012). Outre le génome humain, le génome de plusieurs autres espèces tout comme des banques de données plus spécialisées ont été examinées pour la présence de G4P (Rawal et al., 2006; Hershman et al., 2008; Verma et al., 2008). Chez l'humain, ces analyses ont démontré un enrichissement en G4P dans les régions promotrices, les télomères, les éléments de régulation, les sites de recombinaison, les régions libres de nucléosomes ainsi que les introns et les 5'- et 3'-"*untranslated regions*" (UTR) des ARN messagers (ARNm)(Eddy and Maizels, 2008; Huppert et al., 2008; Qin and Hurley, 2008; Du et al., 2009; Mani et al., 2009; Halder et al., 2009a; Beaudoin and Perreault, 2010; Xu, 2011). De plus, des corrélations importantes ont été observées entre la présence de G4P dans un gène et sa fonction. Par exemple, les proto-oncogènes semblent être associés à la présence de plusieurs G4P tandis que les gènes suppresseurs de tumeur semblent en être appauvris (Eddy and Maizels, 2006). Ces corrélations et enrichissements dans certaines régions spécialisées du génome suggèrent que les G-quadruplexes sont impliqués dans plusieurs mécanismes et phénomènes cellulaires.

5. Rôles biologiques des G-quadruplexes d'ADN

Les G-quadruplexes d'ADN ont été les premiers à être intensément étudiés dans un contexte de biologie cellulaire. Ces études débutèrent au début des années 1990 alors que celles portées sur les G-quadruplexes d'ARN prirent leur envol seulement au milieu des années 2000 comme nous le verrons sous peu. Pour ces raisons, les processus et mécanismes biologiques impliquant les G-quadruplexes d'ADN sont encore aujourd'hui les plus connus et les mieux définis.

Un des rôles pionniers attribué à cette structure d'ADN est son implication dans la maintenance et le cycle de vie des télomères (Intro, Figure 8C)(Williamson et al., 1989; Xu, 2011). Les télomères sont des répétitions riches en guanosines retrouvées aux extrémités des chromosomes. Leur rôle principal est de maintenir la stabilité et l'intégrité du génome au cours des différents cycles de divisions cellulaires. En fait, toutes les séquences télomériques eucaryotes, à l'exception des levures, sont capables de former des G-quadruplexes *in vitro* (Tran et al., 2011). Il semble que ce soit une caractéristique qui a été conservée parmi toutes les espèces, de l'humain aux insectes. L'une des preuves les plus élégantes de la formation de G-quadruplexe *in cellulo* provient de la production d'anticorps contre les structures G-quadruplexes formées par la séquence télomérique du cilié *Stylonychia lemanae* (Schaffitzel et al., 2001). Dans cet article, des expériences d'immunofluorescence sur des noyaux de *Stylonychia lemanae* semblent démontrer la formation de G-quadruplexes au niveau des télomères de cet organisme. La structure G-quadrulplexe formée au niveau des télomères semble avoir un rôle important dans le recrutement et l'orchestration des facteurs impliqués dans l'élongation des télomères et la protection du génome (Paeschke et al., 2005; 2008; Smith et al., 2011; Vannier et al., 2012). L'incapacité de la télomérase (enzyme responsable de la synthèse et l'élongation des télomères) à utiliser la structure G-quadruplexe comme substrat efficace initia un nouvel engouement, qui sera discuté davantage plus loin, celui de cibler les G-quadruplexes comme nouvelles cibles anti-tumorales.



Intro, Figure 8. Quelques fonctions des G-quadruplexes *in cellulo*.

Formation de G-quadruplexes à l'intérieur du génome lors de processus où la double hélice d'ADN se dénature en deux molécules simples brins, p. ex. lors de la transcription (**A**) et de la réplication (**B**). (**A**) Les G-quadruplexes présents dans les promoteurs sont capables d'inhiber la transcription du gène en aval. (**B**) Les G-quadruplexes formés dans le brin matrice lors de la réplication peuvent entraîner des arrêts de réplication et augmenter l'instabilité génomique. (**C**) Les extensions simples brins du brin riche en guanosines des télomères peuvent adopter certains G-quadruplexes possédant différents rôles dans l'entretien des télomères. (**D**) À l'extérieur du noyau, des G-quadruplexes peuvent se former dans les 5'-UTR des ARNm et agir comme répresseur de la traduction. Figure modifiée de (Lipps and Rhodes, 2009).

Un autre rôle très bien documenté implique les G-quadruplexes présents dans les promoteurs de plusieurs gènes, spécifiquement les proto-oncogènes (Intro, Figure 8A). En 1998, l'élément du promoteur responsable de 75-85% du niveau de transcription du proto-oncogène *c-myc*, appelé "nuclease-hypersensitive element III_1" (NHE III_1), fût démontré pour former un G-quadruplexe *in vitro* (Simonsson et al., 1998). En 2002, le groupe du Dr Laurence Hurley démontra que la structure G-quadruplexe présente dans l'élément NHE III_1 était impliquée dans la régulation de la transcription du gène *c-myc in cellulo* (Siddiqui-Jain et al., 2002).

Par la suite, une pléthore de G-quadruplexes dans les régions promotrices de différents proto-oncogènes ont été identifiés comme des éléments de régulation de la transcription (*c-kit*, *VEGF*, *HIF-1α*, *bcl-2*, *k-ras*, etc.)(De Armond et al., 2005; Rankin et al., 2005; Sun et al., 2005; Cogoi and Xodo, 2006; Dai et al., 2006). Dans tous les cas, les structures G-quadruplexes semblent agir comme répresseurs transcriptionnels et différents mécanismes ont été proposés concernant leur fonctionnement (Qin and Hurley, 2008). De plus, certaines approches à la grandeur du génome de la levure et de l'humain ont été réalisées afin d'étudier davantage ce phénomène (Hershman et al., 2008; Verma et al., 2009). Une corrélation importante a été observée entre ce rôle des G-quadruplexes existants dans les promoteurs et le phénotype cancéreux (Verma et al., 2009).

Finalement, d'autres régions du génome semblent être capables de former des structures G-quadruplexes dont la région codant pour les gènes ribosomaux, les régions susceptibles à la recombinaison, les régions de pauses de la réplication et de la transcription (Intro, Figure 8B), certains micro-satellites, etc (Duquette et al., 2004; Larson et al., 2005; Hershman et al., 2008; Piazza et al., 2010; Broxson et al., 2011; Lopes et al., 2011; Paeschke et al., 2011). Le rôle précis de ces structures dans ces phénomènes reste souvent à être investigué davantage. Toutefois, ces études sont vraiment d'une importance capitale, car elles ont permis de confirmer la formation de G-quadruplexes dans un contexte *in cellulo*. Sachant ceci, le champ d'étude des G-quadruplexes présent dans la cellule a réellement pu prendre son envol et bientôt s'élargir encore davantage, se tournant également vers les G-quadruplexes retrouvés dans le transcriptome.

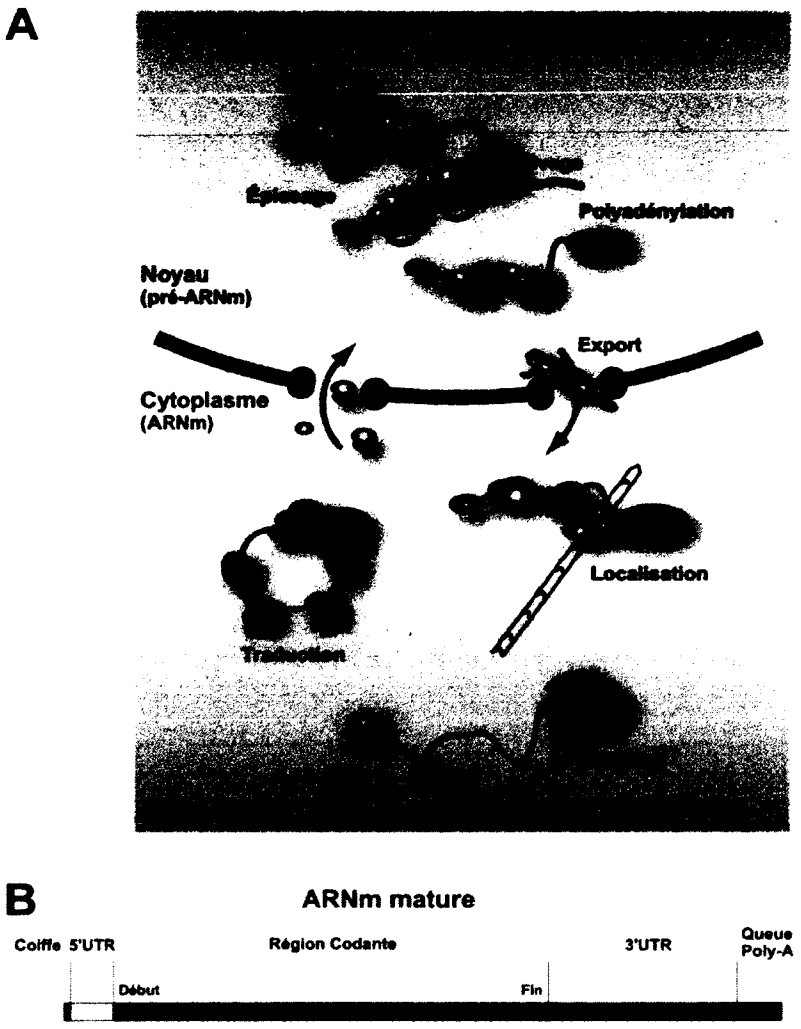
6. Du génome au transcriptome

En 2001, une série de 2.91 milliards de lettres A, T, G et C représentant le secret de la vie humaine fût dévoilée à tous (Lander et al., 2001; Venter et al., 2001). Le projet intitulé "*The Human Genome Project*", impliquant plusieurs groupes de

recherche de partout à travers le monde, révéla pour la première fois l'ampleur du génome humain. Cette carte génomique a été mise à jour en 2003 et ce projet d'envergure titanesque a été achevé officiellement en 2004 (International Human Genome Sequencing Consortium, 2004). À cette époque, ce projet permit d'évaluer le nombre de gènes du génome humain à 20,000-25,000 et que l'information permettant la synthèse de protéines, les exons, couvrait un très maigre 1.1% du génome. Cette évaluation estimait également que 75% du génome correspondait à de l'ADN intergénique. Au même moment où ce premier projet colossal prit fin, un nouveau, tout aussi grandiose, débuta. Le projet multidisciplinaire et international intitulé ENCODE ("*ENCyclopedia Of DNA elements*") fût lancé en septembre 2003. Le but de ce projet est d'identifier tous les éléments fonctionnels du génome humain (ENCODE Project Consortium, 2004). En 2007, les résultats préliminaires de ce projet furent dévoilés et chamboulèrent complètement le monde scientifique (Birney et al., 2007). Ils démontrèrent qu'en réalité la quasi totalité (>90%) du génome est activement transcrit. Il y a seulement quelques années encore, les livres de références parlaient d'environ 5% du génome qui était transcrit et, soudainement, cette proportion augmenta à plus de 90%. Le transcriptome cellulaire devenant immensément plus grand et complexe. Plusieurs nouvelles catégories d'ARN apparurent ainsi qu'une multitude de nouvelles fonctions pour l'ARN (Rinn and Chang, 2012; Sana et al., 2012). Malgré ceci, le nombre total de gènes codants pour des protéines reste sensiblement le même. En effet, la majorité de ces nouveaux transcrits ne codent pour aucune protéine. Ils utilisent plutôt leurs nouvelles fonctions afin d'avoir un impact direct sur le niveau d'expression génique des 20,000-25,000 gènes cellulaires. La découverte de plusieurs de ces nouveaux mécanismes de régulation fait en sorte que la synthèse protéique à partir d'un gène est maintenant plus complexe que jamais (Sana et al., 2012). Toutefois, une chose demeure, les ARNm y jouent toujours un rôle primordial.

6.1 L'ARNm

Le cycle de vie d'un ARNm est parsemé d'événements de maturation qui sont eux-mêmes soumis à une régulation très fine au niveau post-transcriptionnel (Intro, Figure 9A)(McKee and Silver, 2007; Dahan et al., 2011). Chaque gène codant pour une protéine est tout d'abord transcrit en ARN pré-messager (pré-ARNm) qui est le précurseur de l'ARNm. En même temps qu'il est transcrit, ce pré-ARNm subit déjà ses premiers événements de maturation. Peu de temps après le début de la transcription, une structure coiffe est ajoutée en 5'. Cette structure protège l'ARNm des exonucléases cellulaires 5'-3' et favorise la traduction. Toujours de manière co-transcriptionnelle, le pré-ARNm contenant exons et introns est soumis au processus d'épissage qui a comme fonction d'exciser les introns et de positionner bout à bout les exons qui constitueront la séquence de l'ARNm mature. La polyadénylation est un autre processus de maturation très important survenant à l'extrémité 3' de l'ARNm. Il consiste au clivage du pré-ARNm, à un endroit précis déterminé par une série d'éléments de régulation en *cis*, et à l'ajout d'une queue de poly-adénosines (poly-A) qui n'est pas encodée dans le génome d'ADN (Intro, Figure 9A). Cette queue poly-A augmente la stabilité de l'ARNm mature, le protégeant des exonucléases 3'-5', ainsi que le niveau de traduction en interagissant avec la structure coiffe en 5' via un complexe protéique. Le processus de polyadenylation est intimement lié avec les mécanismes de terminaison de la transcription. Cette dernière se terminant quelques nucléotides en aval du site de polyadénylation. La queue poly-A permet également à l'ARNm mature d'être proprement exporté au cytoplasme où il pourra finalement agir comme matrice d'information et, avec l'aide des ribosomes et des ARN de transfert (ARNt), mener à la synthèse de protéines.



Intro, Figure 9. Cyle de vie des ARNm.

(**A**) Les ARNm peuvent être régulés à chacune des étapes de leur cycle de vie, soit par la liaison de différentes protéines ou par la présence de plusieurs éléments de régulation post-transcriptionnelle. À l'intérieur du noyau, le pré-ARNm est transcrit, coiffé, épissé, clivé et polyadénylé de manières co- et post-transcriptionnelles. Suite à un contrôle de qualité de l'ARNm mature ainsi produit, seuls les transcrits proprement maturés sont exportés vers le cytoplasme. Une fois dans le cytoplasme, l'ARNm est soumis à différents mécanismes incluant la localisation subcellulaire, la traduction et la dégradation. Ceux-ci seront déterminés par la multitude d'éléments de régulation post-transcriptionnelle présents à l'intérieur de chaque ARNm et auront un impact majeur sur les niveaux de synthèse protéique tout comme sur la composition du transcriptome cellulaire. (**B**) Schéma des parties principales de l'ARNm mature, incluant la structure coiffe, les régions 5'- et 3'-UTR, la région codante et la queue poly-A. Figure adaptée de (McKee and Silver, 2007).

Cette synthèse protéique appelée traduction est étroitement liée à la stabilité et au "temps de demi-vie" de l'ARNm. C'est pourquoi le processus de dégradation de l'ARNm est aussi soumis à divers moyens de régulation. La traduction s'effectuant majoritairement au cytoplasme et au réticulum endoplasmique rugueux, le transport de l'ARNm à ces zones de traduction actives est également très important. L'ARNm mature ainsi produit comprend différentes régions importantes (Intro, Figure 9B). À l'extrémité 5' on retrouve une structure coiffe suivie d'une région 5' non traduite (5'-UTR pour "*UnTranslated Region*"), d'une région codante; composée d'une série de triplets de nucléotides (codons) dictant la séquence des acides aminés à être incorporés dans la protéine à synthétiser, d'une région 3' non traduite (3'-UTR) et d'une queue poly-A. Si les rôles joués par la structure coiffe, la région codante et la queue poly-A semblent relativement clairs et bien définis, ceux attribués aux 5'- et 3'-UTR sont plus divers et en évolution constante. Toutefois, un thème porteur en émerge: ces deux régions de l'ARNm sont particulièrement riches en éléments de régulation post-transcriptionnelle capables d'influencer la plupart des étapes du cycle de vie de l'ARNm.

6.2 Les éléments de régulation post-transcriptionnelle et l'ARNm

La régulation du niveau de la transcription a longtemps été perçue comme le moyen primaire de réguler l'expression génique et l'accumulation de transcrits. Toutefois, avec ce transcriptome cellulaire immensément plus grand que prévu, il est clair que les éléments de régulation post-transcriptionnelle deviennent les pierres angulaires déterminant le destin de chaque molécule d'ARN transcrite. Ils doivent avoir évolués de façon à permettre une accumulation dynamique de certains sous-groupes de transcrits, à l'intérieur de ce vaste transcriptome, variant en fonction des différents stimuli perçus par la cellule. Ces éléments peuvent correspondre à des séquences primaires de l'ARN, généralement des sites de liaison reconnus par des protéines spécifiques, ou à des structures secondaires et tertiaires de l'ARN. L'importance de la structure secondaire et tertiaire de l'ARN pour la fonction de plusieurs éléments de régulation post-transcriptionnelle est maintenant bien établie et soulève même un engouement dans le but de mieux

comprendre ces relations structures-fonctions. Cet enthousiasme se caractérise bien avec le désir de plusieurs scientifiques de décoder le structurome de l'ARN ("*RNA structurome*")(Wan et al., 2011). Avec l'émergence de techniques de séquençage de plus en plus puissantes, l'idée d'étudier la structure adoptée par tous les ARN cellulaires en une seule expérience n'est plus farfelue (Westhof and Romby, 2010).

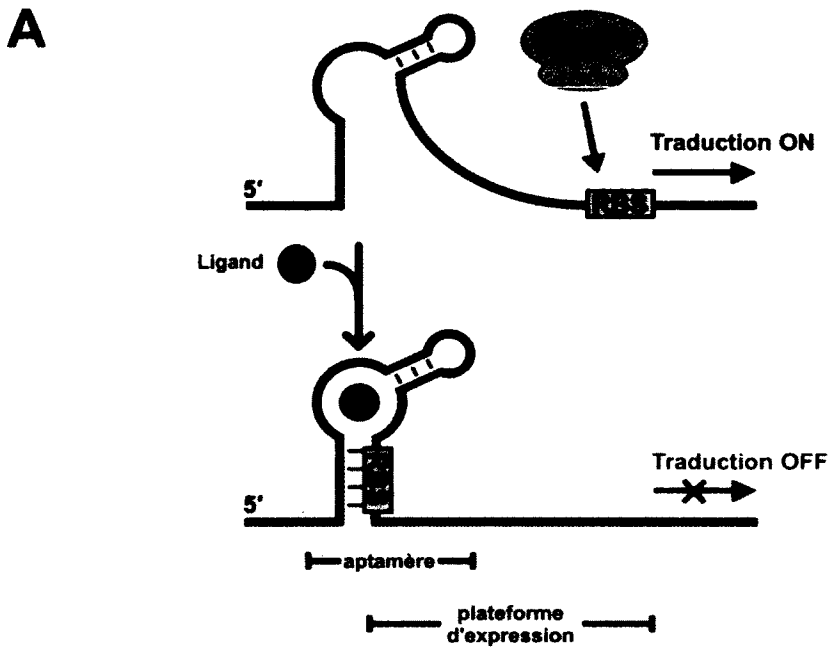
N'échappant pas à cette règle, l'ARNm s'avère être un chef de file en la matière. En effet, l'ARNm est parsemé d'une panoplie d'éléments de régulation post-transcriptionnelle capables d'influencer la totalité des étapes de sa maturation et de son cycle de vie, affectant ultimement le niveau de protéines synthétisées par celui-ci (Moore, 2005; Licatalosi and Darnell, 2010). Comme il a été mentionné plus haut, les régions 5'- et 3'-UTR sont particulièrement riches en de tels éléments (Barrett et al., 2012).

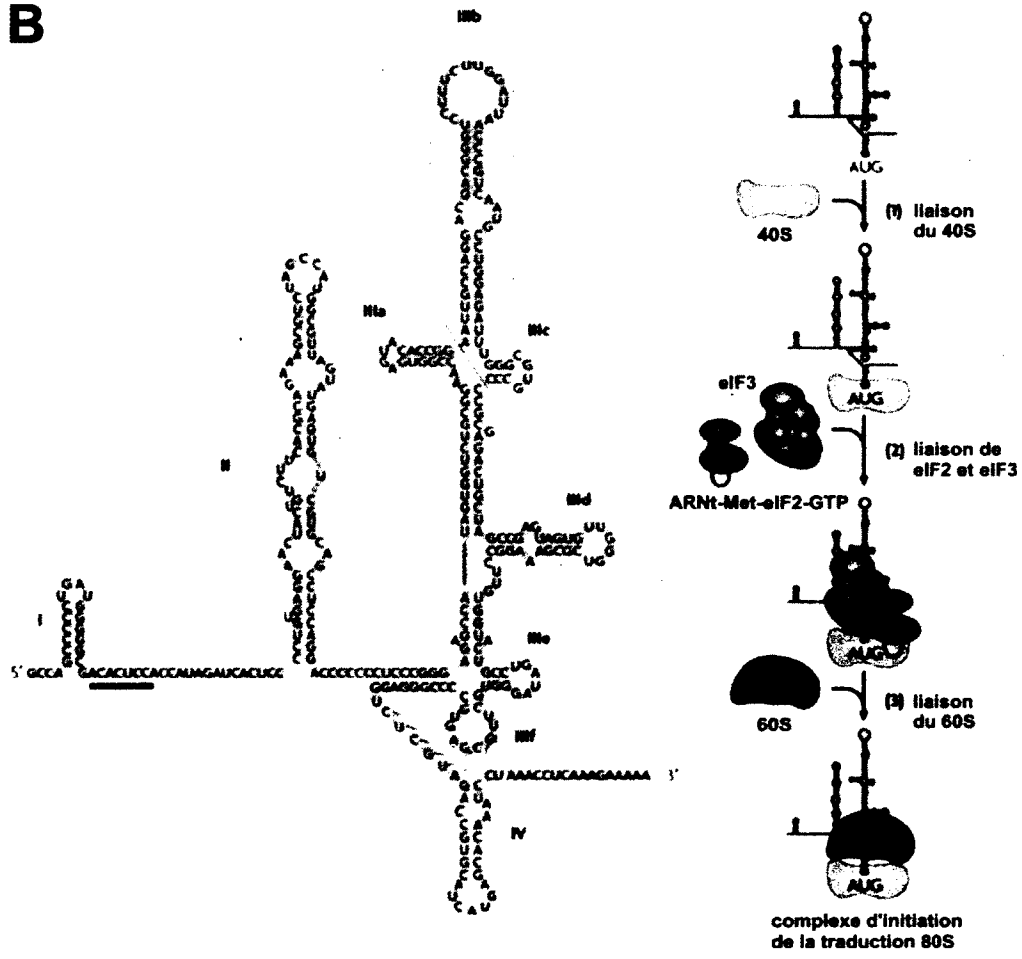
6.2.1 Éléments post-transcriptionnels dans les 5'-UTR

La région 5'-UTR des ARNm contient plusieurs éléments post-transcriptionnels. La plupart de ceux-ci sont impliqués dans la régulation du niveau de traduction. Parmi les plus connus on retrouve sans aucun doute les fameux riborégulateurs (ou "*riboswitches*"). Les riborégulateurs sont littéralement des senseurs d'ARN capables de détecter des changements de stimuli cellulaires en absence de cofacteurs protéiques (Henkin, 2008). Un riborégulateur est typiquement composé de deux domaines: un domaine appelé aptamère qui est capable de lier de façon spécifique un ligand particulier et une plateforme d'expression qui est capable, selon sa structure, d'affecter l'expression génique (Intro, Figure 10A). Lorsque le ligand se lie de façon spécifique à l'aptamère, le riborégulateur, incluant la plateforme d'expression, subit un changement au niveau de sa structure secondaire se reflétant par un changement au niveau de l'expression du gène. Un type de plateforme fréquemment retrouvé est celui qui, suite à la liaison du ligand à l'aptamère, change de conformation afin de former une tige boucle stable

séquestrant le site d'entrée des ribosomes (RBS pour "*ribosome-binding site*"), réduisant de beaucoup la traduction de son ARNm (Intro, Figure 10A)(Breaker, 2012). Un autre exemple d'élément post-transcriptionnel est l'ARNm codant pour la protéine précurseur de l'amyloïde (APP) qui contient dans son 5'-UTR un élément de réponse au fer (IRE). Cet élément est une structure d'ARN tige-boucle capable de lier la protéine régulatrice du fer 1 (IRP1) de manière dépendante du fer. En absence de fer, la protéine IRP1 possède une grande affinité pour l'IRE et se lie de façon efficace à l'ARNm de APP. Cette liaison efficace entraîne une inhibition de la traduction et une réduction de la synthèse protéique. En présence de fer, la protéine IRP1 perd son affinité pour l'IRE et ne peut plus lier efficacement l'ARNm de APP. Cette perte d'affinité occasionne une augmentation du niveau de traduction et, par le fait même, une plus grande accumulation de la protéine APP dans la cellule (Cho et al., 2010).

Il existe aussi des structures d'ARN très complexes au niveau de leur structure secondaire capable de promouvoir la traduction d'un ARNm. Ces structures sont appelées des sites d'entrée internes pour les ribosomes (IRES pour "*internal ribosome entry sites*")(Intro, Figure 10B)(Jackson, 2005). Le processus de traduction est généralement dépendant de la structure coiffe des ARNm. Toutefois, il existe certains ARNm capables d'initier la traduction de façon indépendante de la structure coiffe via l'utilisation d'un IRES. L'IRES est typiquement formé d'un agencement complexe de structures secondaires et tertiaires d'ARN. Ce motif permet le recrutement et l'assemblage du complexe d'initiation de la traduction, normalement recruté au niveau de la structure coiffe (Intro, Figure 10B)(Fraser and Doudna, 2007). Les IRES ont initialement été découverts chez les virus, le plus étudié étant certainement celui présent dans le génome du virus de l'hépatite C (VHC). Outre les virus, plusieurs ARNm cellulaires possèdent également une structure IRES dans leur 5'-UTR (p. ex. bcl-2, apaf-1, c-myc, p53 et cyclin D1)(Pichon et al., 2012). Ces ARNm peuvent donc continuer à être traduit de manière efficace même lorsque le mécanisme de traduction dépendant de la structure coiffe est inhibé.


A
Traduction ON
5'
RBS
Ligand
Traduction OFF
5'
aptamère
plateforme
d'expression


B
IIIb
IIIa
IIIc
II
IIId
I
IIIe
IIIf
IV
AUG
40S
(1) liaison
du 40S
eIF3
AUG
(2) liaison de
eIF2 et eIF3
ARNt-Met-eIF2-GTP
AUG
60S
(3) liaison
du 60S
AUG
complexe d'initiation
de la traduction 80S

Intro, Figure 10. Riborégulateur et IRES.

(**A**) Régulation de la traduction par un riborégulateur. En absence de ligand, le ribosome peut se lier au RBS ("*ribosome-binding site*") de l'ARNm et initier la traduction. Quand le ligand est disponible et se lie à l'aptamère, le RBS est séquestré à la suite d'un changement de structure de l'ARN et ne peut plus être reconnu par le ribosome, inhibant ainsi la traduction de l'ARNm. (**B**) (côté gauche) Représentation du 5'-UTR du génome du virus de l'hépatite C. La structure secondaire du IRES est composée de trois domaines (II-IV) et différents sous-domaines (a-f) pour le domaine III. Le codon d'initiation AUG est coloré en rouge. (côté droit) Recrutement des différents facteurs d'initiation de la traduction par les différentes structures du IRES. Figure adapté de (Kim and Breaker, 2008) et (Fraser and Doudna, 2007).

6.2.1 Éléments post-transcriptionnels dans les 3'-UTR

Les éléments post-transcriptionnels présents dans les 3'-UTR peuvent réguler plusieurs processus incluant le clivage de l'ARNm, sa stabilité, sa polyadénylation, sa traduction et sa localisation. L'une des régulations les plus importantes s'effectuant majoritairement au niveau des 3'-UTR est celle dictée par les microARNs (miARN)(Winter et al., 2009; Winter and Diederichs, 2011). Les miARN sont des ARN non codants subissant plusieurs étapes de maturation menant ultimement à la synthèse de leur forme mature d'une longueur moyenne de 22 nucléotides. Ces miARN matures s'associent à certaines protéines afin de former le complexe RISC ("*RNA-induced silencing complex*") capable d'interagir avec les ARNm de façon post-transcriptionnelle et de réguler leur expression (Fabian et al., 2010). Chez les mammifères, le complexe RISC exerce généralement son effet via un appariement partiel de paires de bases du miARN à un site de liaison présent dans la séquence de l'ARNm cible (généralement dans le 3'-UTR). Le miARN chargé dans le complexe RISC reconnaît son site de liaison majoritairement avec l'aide de sa région "*seed*" (nucléotides des positions 2 à 8 du miARN) qui se doit d'être parfaitement appariée à l'ARNm, alors que le reste de sa séquence peut ne posséder qu'une très faible complémentarité avec celui-ci (Bartel, 2009). Typiquement, la liaison du complexe RISC à un ARNm entraine une diminution importante de son niveau de traduction. Il existe une panoplie de gènes encodant pour des miARN dans le génome humain. Aujourd'hui, une grande quantité

d'information est maintenant disponible concernant l'expression et la fonction des miARN. Il est maintenant évident qu'ils sont une composante vitale du contrôle de l'expression génique étant activement impliqué dans plusieurs des événements biologiques les plus importants incluant: la prolifération et la différentiation cellulaire, le développement, la régulation du système nerveux et la tumorigenèse (Huang et al., 2011).

D'autres éléments de régulation présents dans les 3'-UTR sont les régions riches en adénosines (A) et uraciles (U) (ARE pour "*AU-rich element*"). Elles ont entre 50 et 150 nucléotides de long, elles possèdent beaucoup de A et U (souvent de multiples répétitions du pentanucléotide AUUUA) et elles sont pratiquement exclusivement retrouvées dans la partie 3'-UTR des ARNm (Chen and Shyu, 1995). Les ARE sont reconnues par une grande variété de protéines. Une fois le complexe ARE-protéine formé, la stabilité de l'ARNm est généralement grandement diminuée suite à une augmentation de son niveau de dégradation. Promouvoir la dégradation d'un ARNm est un bon moyen d'abaisser son accumulation dans la cellule et, du même coup, diminuer la quantité de protéine synthétisée par ce gène (Elkon et al., 2010).

Bien que ces deux derniers mécanismes de régulation génique soient basés majoritairement sur la séquence primaire de l'ARN, la présence de structures secondaires spécifiques a déjà été démontrée comme étant capable de réguler chacun de ces deux phénomènes. En effet, il peut arriver qu'un site de miARN soit séquestré dans une structure tige boucle rendant son association avec le complexe RISC impossible. C'est le cas pour le site de liaison du miARN-221/222 dans le 3'-UTR de l'ARNm du gène p27 (Kedde et al., 2010). Ce site de liaison forme une structure tige boucle en s'appariant à une séquence correspondant au site de liaison de la protéine Pumilio-1 (PUM1). Dans ce cas, la liaison de PUM1 à son site de liaison est nécessaire afin de libérer le site de liaison du miARN et de permettre à celui-ci d'être reconnu par le complexe RISC et de réguler l'expression

de l'ARNm. De plus, une autre structure tige boucle est capable de moduler quelle protéine aura une meilleure affinité pour une ARE particulière (Fialcowitz et al., 2005). L'ARNm du gène encodant le facteur de nécrose tumorale alpha (TNFα) contient une ARE dans son 3'-UTR qui est capable d'adopter une structure tige boucle. En absence de la formation de cette structure secondaire, la ARE possède une haute affinité pour la protéine $p37^{AUF1}$ ce qui mène à une dégradation rapide de l'ARNm. Inversement, si la structure tige boucle est formée, la ARE perd son affinité pour $p37^{AUF1}$ et est capable de mieux lier la protéine de choc thermique Hsp70 augmentant la stabilité de l'ARNm et une meilleure expression génique (Fialcowitz et al., 2005). Indubitablement, les différentes fonctions de l'ARNm peuvent être régulées autant par la séquence primaire que la structure secondaire de différents éléments de régulation post-transcriptionnelle présents dans leurs régions 5'- et 3'-UTR. La séquence en acide nucléique et la structure secondaire ou tertiaire adoptée par l'ARN travaillent souvent de concert afin de réguler l'expression génique. Les G-quadruplexes d'ARN retrouvés dans le transcriptome n'échappent pas à cette règle. Ils commencent également, petit à petit, à faire ressentir leur présence.

6.3 Les G-quadruplexes dans le transcriptome cellulaire

Comme mentionné précédemment, il semble clair que les G-quadruplexes d'ADN ont un impact sur plusieurs mécanismes cellulaires. Or, le génome d'ADN cellulaire est majoritairement sous une conformation de double hélice de type B, où chacun des brins s'apparie via des paires de bases Watson-Crick afin de former un duplex. L'ARN, sans brin complémentaire le forçant à adopter une structure double hélice, peut adopter une pléiade de structures. De ce fait, l'ARN riche en guanosines serait plus favorable à la formation d'un G-quadruplexe que l'ADN. Ajouter à cela que pour une même séquence formant un G-quadruplexe, la structure en ARN est typiquement plus stable que celle à base d'ADN (Saccà et al., 2005). Il est donc fortement probable que l'on retrouve à l'intérieur du transcriptome plusieurs G-quadruplexes et que ceux-ci soient capables d'influencer différents mécanismes et

phénomènes cellulaires. Les découvertes des cinq dernières années démontrent bien que c'est le cas (pour une revue voir (Ji et al., 2011; Millevoi et al., 2012)).

Dans la première année de mes études graduées, en 2007, une étude pionnière concernant l'étude des G-quadruplexes d'ARN biologiques rapporta qu'un G-quadruplexe présent dans un 5'-UTR d'un ARNm pouvait agir comme répresseur traductionnel *in vitro* (Kumari et al., 2007). Selon le mécanisme d'action proposé, cette structure stable agirait comme bloc stérique provoquant le décrochement de la petite sous-unité ribosomale lorsque cette dernière, recrutée via la structure coiffe, est à la recherche du codon d'initiation de la traduction AUG (Intro, Figure 8D). Suivant cette preuve de concept réalisée en extrait de réticulocytes, plusieurs autres études (incluant une présentée dans cette thèse) ont rapporté ce phénomène dans plusieurs ARNm différents, ce qui en fait sans aucun doute le rôle pour les G-quadruplexes d'ARN biologiques le plus documenté aujourd'hui. Une revue portant principalement sur le sujet a tout récemment été publiée (Bugaut and Balasubramanian, 2012). Toutefois, un G-quadruplexe présent dans un 5'-UTR n'agit pas toujours comme répresseur traductionnel. Au contraire, il peut même permettre une initiation de la traduction efficace. En effet, il a été rapporté qu'une structure G-quadruplexe peut faire partie d'un élément IRES, comme c'est le cas pour l'ARNm codant pour le facteur de croissance des fibroblastes 2 (FGF-2) (Bonnal et al., 2003). Ce dernier possède une activité IRES au niveau de son 5'-UTR. Suite à la caractérisation de ce IRES, il s'avère qu'il contient une structure G-quadruplexe et que celle-ci est importante afin de permettre une bonne initiation de la traduction.

La présence des G-quadruplexes d'ARN ne se limite pas aux 5'-UTR des ARNm, ils se retrouvent à plusieurs endroits différents, incluant les introns. Il a été démontré qu'une structure G-quadruplexe présente dans l'intron 3 du gène *TP53* (encodant pour la protéine p53) régule l'épissage alternatif de son pré-ARNm au niveau de l'intron 2 (Marcel et al., 2011). La formation du G-quadruplexe

favoriserait l'excision adéquate de l'intron 2 et la production d'une protéine p53 active. En plus de caractériser ce phénomène via une approche structurale, mutationnelle et phénotypique, les auteurs ont également vérifié l'impact de la présence d'un ligand, liant et stabilisant spécifiquement les structures G-quadruplexes, sur celui-ci. La présence du ligand menait à une amplification du niveau d'épissage promu par la formation du G-quadruplexe, supportant fortement leur hypothèse. Toujours pour le gène *TP53*, il a tout récemment été démontré qu'un G-quadruplexe, présent dans son pré-ARNm et situé en aval du site de polyadénylation, était primordial afin de maintenir un bon niveau de polyadénylation suite à un stress cellulaire causant des dommages au niveau du génome d'ADN (Decorsière et al., 2011). Plus précisément, la présence de dommages dans l'ADN cellulaire mène à une augmentation de la production de protéines hnRNP H/F. Ces protéines fraichement synthétisées lient le pré-ARNm du gène *TP53* via le G-quadruplexe situé en aval du site de polyadénylation et favorisent le recrutement efficace des facteurs généraux de polyadenylation permettant un niveau de polyadenylation adéquat. Ce mécanisme de régulation permet de maintenir une expression appropriée de la protéine p53, qui possède un rôle majeur dans le processus de réparation de l'ADN.

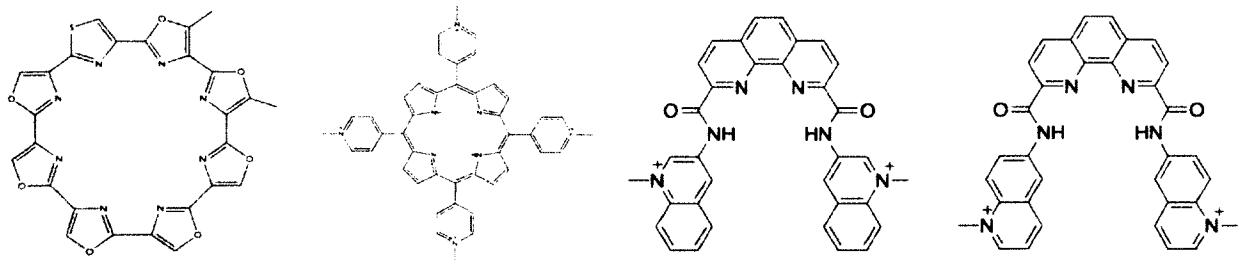
Finalement, une étude récente rapporte qu'une structure G-quadruplexe est présente dans les 3'-UTR des ARNm codants pour deux protéines post-synaptiques importantes, PSD-95 et CaMKIIa (Subramanian et al., 2011). En plus d'être présent dans ces deux ARNm, ces G-quadruplexes agissent comme signaux de localisation aux neurites des cellules neuronales. Il est suggéré que ces structures pourraient servir de site de liaison à la protéine FMRP ("*Fragile X Mental Retardation Protein*") et que cette association permettrait le transport de l'ARNm au bon endroit dans la cellule. FMRP est probablement la protéine possédant une affinité pour les G-quadruplexes d'ARN la plus documentée (pour une revue voir (Melko and Bardoni, 2010; Sissi et al., 2011)). Selon des résultats d'une analyse de structure à haute résolution réalisée par RMN, elle reconnaitrait la jonction entre un duplex et un G-quadruplexe d'ARN (Phan et al., 2011). Elle se

lierait à des régions précises d'une multitude de transcrits possédant une forte probabilité à former des G-quadruplexes (Darnell et al., 2001). La perte de fonction de FMRP entraine une dérégulation au niveau traductionnel de plusieurs de ces ARNm (Brown et al., 2001). En conclusion, il semble évident que les G-quadruplexes d'ARN présents dans le transcriptome cellulaire peuvent avoir un impact important et déterminant à l'égard de plusieurs mécanismes et phénomènes.

7. Les G-quadruplexes comme cibles thérapeutiques

Comme il a été souligné, les G-quadruplexes sont impliqués dans plusieurs processus et mécanismes cellulaires. Les plus connus d'entre eux (réliés aux télomères et aux promoteurs de proto-oncogènes) sont directement associés à des joueurs clés dans le développement de cancers. Il n'en fallait pas plus pour lancer une réelle frénésie dans la recherche de petites molécules capables de se lier spécifiquement à la structure G-quadruplexe et d'affecter certains de ces différents processus cellulaires (Huppert, 2007; Neidle and Parkinson, 2008; Collie and Parkinson, 2011; Xu, 2011; Düchler, 2012). Le premier grand défi était de trouver des ligands capables de discriminer un duplex d'un tétraplex d'ADN. Pour ce faire, les chercheurs se sont tournés vers l'une des caractéristiques structurales la plus évidente concernant les G-quadruplexes, leur très grande surface d'interactions de type π (environ le double du duplex d'ADN). En mettant cette caractéristique à profit, il a été possible de synthétiser certains ligands possédant une spécificité pour la structure G-quadruplexe d'ADN jusqu'à 10,000 fois plus élevée que pour le duplex (Dixon et al., 2007). La première évidence qu'une petite molécule pouvait inhiber l'activité de la télomérase via la stabilisation de la structure G-quadruplexe a été démontrée en 1997. Un composé dérivé de l'anthraquinone était capable d'inhiber l'activité de la télomérase dans un essai TRAP ("*Telomeric Repeat Amplification Protocol*")(Sun et al., 1997). Depuis, plusieurs autres ligands ont été développés et étudiés dans cette optique (Xu, 2011). Ils partagent typiquement trois caractéristiques communes: i) un grand centre aromatique maximisant les

interactions de type π avec la surface des G-quartets; ii) une charge positive afin d'interagir avec les charges négatives du squelette phosphate de l'ADN; et, iii) diverses chaînes latérales afin d'optimiser l'introduction de groupements fonctionnels spécifiques interagissant avec les sillons et les boucles du G-quadruplexe (Intro, Figure 11). Parmi ceux-ci on retrouve un composé naturel retrouvé chez l'organisme *Streptomyces anulatus*, la télomestatine (Intro, Figure 11). Ce dernier est le composé le plus efficace à lier la structure G-quadruplexe connu à ce jour et possède une constante d'inhibition de la télomérase de 5 nanomolaire (nM) (Kim et al., 2002). L'existence de G-quadruplexes, à l'intérieur de plusieurs promoteurs de proto-oncogènes, importants dans la régulation de leur transcription et de leur niveau d'expression permit l'apparition d'une nouvelle série de cibles antitumorales. Un exemple précoce est l'utilisation de la molécule $TMPyP_4$, un dérivé porphyrique, afin de cibler et stabiliser une structure G-quadruplexe présente dans la région promotrice du proto-oncogène *c-myc* menant à une diminution d'environ 80% de son niveau de transcription (Intro, Figure 11)(Siddiqui-Jain et al., 2002). Des résultats similaires impliquant d'autres molécules et/ou d'autres proto-oncogènes ont également largement été rapportés (Intro, Figure 11)(Qin and Hurley, 2008).



Intro, Figure 11. La structure moléculaire de certains ligands spécifiques pour les G-quadruplexes.

Figure adaptée de (Huppert, 2008) et (Bugaut and Balasubramanian, 2012).

Sachant que la télomérase est l'enzyme qui permet l'extension des télomères et qu'une augmentation de son activité est quasiment un incontournable pour la survie de toute cellule cancéreuse, tout comme une augmentation de l'expression de certains proto-oncogènes comme *c-myc*, les espoirs que les G-quadruplexes soient des cibles anticancéreuses efficaces sont prometteurs. Toutefois, bien qu'il soit présentement possible de faire la discrimination entre un duplex et un G-quadruplexe avec l'aide d'un composé chimique, le défi de faire la différentiation entre différentes topologies de G-quadruplexes est encore bien réel et loin d'être achevé. En effet, avec plus de 376 000 G4P dans le génome humain, la nécessité de développer des ligands ciblant un spectre étroit de topologies est cruciale pour le futur. Dans le même ordre d'idée, il sera important de synthétiser des molécules capables de différencier les G-quadruplexes d'ADN et d'ARN. Pour l'instant, il existe certains ligands capables de lier et de stabiliser des G-quadruplexes d'ARN et d'affecter les mécanismes dans lesquels ils sont impliqués (p. ex. l'épissage et la répression de la traduction)(Gomez et al., 2004; Halder et al., 2011). Cependant, ces molécules ont tout d'abord été synthétisées dans le but de cibler des G-quadruplexes d'ADN et c'est généralement un heureux hasard s'ils possèdent également une bonne affinité pour ceux à base d'ARN. Ceci n'est pas très surprenant puisque les G-quadruplexes d'ADN et d'ARN peuvent partager plusieurs similarités structurales, particulièrement au niveau de leur corps central constitué d'empilement de G-quartets. En résumé, même si plusieurs avancées et preuves de concept intéressantes ont été réalisées dans le but d'utiliser les structures G-quadruplexes comme cibles thérapeutiques, le grand défi qui attend les scientifiques est celui d'être capable de ne cibler qu'un ou qu'une petite poignée de G-quadruplexes parmi la kyrielle présente dans la cellule. Pour ce faire, ces ligands devront probablement reposer sur la reconnaissance des sillons, des boucles, des nucléotides adjacents ou de l'orientation des brins de la structure G-quadruplexe.

8. Hypothèse de recherche

Les propriétés intrinsèques de la structure G-quadruplexe d'ARN suggèrent qu'elle serait idéale pour agir comme élément de régulation. Tout d'abord, elle possède une grande stabilité lui permettant de rivaliser et/ou d'interagir adéquatement avec d'autres éléments hautement structurés. De plus, le contexte cellulaire est très favorable à la formation de G-quadruplexes (c.-à-d. la concentration de potassium, le pH, les séquences riches en guanines, etc). Finalement, la grande diversité topologique et structurale des G-quadruplexes leur donne la possibilité d'être impliqués dans une multitude de mécanismes et de fonctions à l'intérieur de la cellule. Avec la pléthore de fonctions associées aux molécules d'ARN dont l'activité catalytique, la traduction, la polyadénylation, etc, la structure G-quadruplexe, avec ses propriétés et ses caractéristiques, semble être bien armée afin de pouvoir réguler plusieurs d'entre-elles.

9. Objectif de recherche

Mon objectif principal de recherche était d'étudier l'aptitude de la structure G-quadruplexe à agir comme élément régulateur de l'ARN. Pour ce faire, j'ai initialement étudié la capacité d'un G-quadruplexe à interagir avec une autre structure d'ARN très stable et possédant une activité catalytique, le ribozyme du virus de l'hépatite D. Par la suite, j'ai caractérisé la distribution, la formation et l'impact de la présence de G-quadruplexes dans les 5'- et 3'-UTR des ARNm dans un contexte cellulaire.

9.1 Chapitre 1: Le G-quartzyme

9.1.1 Contexte

Le ribozyme du virus de l'hépatite D (Rz VDH) est un motif d'ARN capable de catalyser une réaction de clivage d'un ARN cible, en *cis* ou en *trans*, uniquement lorsque celui-ci se replie suivant un chemin réactionnel bien particulier. Plusieurs études dans le laboratoire ont permis de faire du sentier de repliement du Rz VDH, l'un des plus élaborés pour une molécule d'ARN (pour une revue sur le sujet, voir Annexe 2.1 et 2.2). Ce ribozyme est aussi bien connu pour posséder un haut niveau de complexité structurale ainsi qu'une grande stabilité (Lévesque et al., 2002). Précédemment dans le laboratoire, une étude de SELEX ("Systematic Evolution of Ligands by EXponential Enrichement") a mené à la sélection d'une séquence d'ARN capable de former une structure G-quadruplexe stable en présence de potassium (voir Annexe 2.3). Nous détenions donc deux motifs d'ARN très stable, l'un relié à une activité catalytique et l'autre replié adéquatement uniquement en présence de potassium. Comment ces deux structures étaient capables d'interagir ensemble? Il fallait le tester pour le savoir.

9.1.2 Objectif général

Générer et caractériser une chimère G-quadruplexe-Ribozyme VHD, baptisée le G-quartzyme.

9.1.3 Objectifs spécifiques

i. Évaluer les nouvelles fonctions du G-quartzyme;

ii. Déterminer les paramètres cinétiques propres au G-quartzyme; et,

iii. Étudier le mécanisme d'action du G-quartzyme.

9.2 Chapitre 2: Les G-quadruplexes dans les 5'-UTR

9.2.1 Contexte

Les exemples de G-quadruplexes d'ARN avec un rôle biologique se faisant très rare, l'étude du groupe du Dr Balasubramanian arriva donc comme une nouveauté dans ce champ d'étude (Kumari et al., 2007). Il démontra *in vitro*, qu'un G-quadruplexe présent dans le 5'-UTR du proto-oncogène *NRAS* agissait comme un répresseur traductionnel. Peu de temps après, un autre groupe rapportait le même phénomène chez la bactérie E. *coli* (Wieland and Hartig, 2007). Certains exemples étaient présents concernant la caractérisation de ce phénomène chez les cellules de mammifères. Toutefois, ils s'avéraient plutôt rares, quelque peu redondants et fragmentaires, empêchant une évaluation globale de ce rôle relié à la structure G-quadurplexe.

9.2.2 Objectif général

Faire une analyse globale de la distribution et de l'impact des G-quadruplexes présents à l'intérieur des 5'-UTR des ARNm du transcriptome humain.

9.2.3 Objectifs spécifiques

i) Développer une méthode robuste avec laquelle approcher mon objectif général;

ii) Étudier la distribution des G-quadruplexes potentiels dans les 5'-UTR humains;

iii) Étudier leur repliement *in vitro* et *in cellulo*;

iv) Évaluer et caractériser leur impact *in cellulo* sur l'expression génique; et,

v) Étudier la présence et l'impact de polymorphisme nucléotidiques (SNP pour "*single-nucleotide polymorphism*") à l'intérieur des G-quadruplexes.

9.3 Chapitre 3: Les G-quadruplexes dans les 3'-UTR

9.3.1 Contexte

Alors que l'étude des G-quadruplexes présents dans les 5'-UTR devient de plus en plus étoffée, celle portée sur ceux retrouvés dans les 3'-UTR des ARNm en est à ces balbutiements, voir inexistante. Récemment, il a été rapporté que des G-quadruplexes dans le 3'-UTR de deux gènes pouvaient agir comme motif de localisation cellulaire aux dentrites (Subramanian et al., 2011). Bien que situé en dehors du 3'-UTR de l'ARNm mature, un autre G-quadruplexe d'ARN en 3' du gène encodant pour la protéine p53 a été rapporté comme ayant un effet positif sur l'efficacité de polyadenylation d'un site de clivage situé en amont de celui-ci (Decorsière et al., 2011). Puisque peu de choses sont connues sur les G-quadruplexes présents dans les 3'-UTR et que la lueur de certains rôles potentiels pour ceux-ci semble percer l'horizon, l'intérêt d'en apprendre davantage à leur sujet devient considérable.

9.3.2 Objectif général

Étudier la distribution et l'impact des G-quadruplexes présents à l'intérieur des 3'-UTR des ARNm du transcriptome humain.

9.3.3 Objectifs spécifiques

i) Développer une méthode robuste avec laquelle approcher mon objectif général;

ii) Étudier la distribution des G-quadruplexes potentiels dans les 3'-UTR humains;

iii) Étudier leur repliement *in vitro* et *in cellulo*; et,

iv) Évaluer et caractériser leur impact *in cellulo* sur l'expression génique.

RÉSULTATS

CHAPITRE 1: Genèse et caractatérisation d'une chimère G-quadruplexe-Ribozyme VDH, le G-quartzyme.

ARTICLE: Potassium ions modulate a G-quadruplex-ribozyme's activity

Jean-Denis Beaudoin and Jean-Pierre Perreault

Article publié dans: *RNA* (**2008**) 14: 1018-1025

AVANT-PROPOS:

J'ai réalisé 100% des expériences rapportées dans cet article. J'ai également monté toutes les figures et participé de façon active à l'écriture du manuscrit.

RÉSUMÉ

Le ribozyme du virus de l'hépatite D se replie en une structure tertiaire très compacte. Toutefois, à l'inverse de d'autres ribozymes, il semble incapable de suivre des sentiers de repliements alternatifs. L'ingénierie moléculaire du ribozyme du virus de l'hépatite D a mené au développement d'un ribozyme possédant une activité endonucléase qui est sous le contrôle d'une structure G-quadruplexe (c.-à-d. un G-quartzyme). Cette nouvelle espèce représente une toute nouvelle classe de ribozyme. Des mutants de ce ribozyme ont été générés dans le but d'expliquer cette modulation de l'activité de coupure dépendante de la présence d'une structure G-quadruplexe. Une caractérisation cinétique du G-quartzyme a été réalisée sous différentes conditions et en tenant compte d'un seul cycle catalytique. Il se trouve à être actif uniquement en présence de cations potassiques qui agissent comme contre-ions dans le positionnement des quatre guanines coplanaires formant l'unité de base de la structure G-quadruplexe. Le G-quartzyme réagit comme un ribozyme allostérique, avec les cations potassiques agissant comme effecteurs positifs avec un coefficient de Hill de 2.9 ± 0.2. Le changement de conformation induit par la présence d'ions potassiques est supporté par la cartographie enzymatique et chimique des structures inactives (*off*) et actives (*on*). Cette étude montre qu'il est possible d'interférer avec la structure compacte du ribozyme du virus de l'hépatite D en y ajoutant une structure inhabituellement stable. À notre connaissance, le G-quartzyme est l'unique ribozyme détenant une activité dépendante de cations monovalent.

ABSTRACT

Hepatitis delta virus ribozyme folds into a tightly packed tertiary structure. However, unlike other ribozymes, it does not appear to be able to follow alternative folding pathways. Molecular engineering of the hepatitis delta virus ribozyme led to the development of a ribozyme possessing an endoribonuclease activity that is under the control of a G-quadruplex structure (i.e. a G-quartzyme). This latter species represents an entirely new class of ribozyme. Mutants of this ribozyme were then generated in order to shed light on the modulation of the cleavage activity caused by the presence of the G-quadruplex structure. Kinetic characterization of the G-quartzyme was performed under various single turnover conditions. It was found to be active only in the presence of potassium cations that act as counter ions in the positioning of the four co-planar guanines that form the building block of the G-quadruplex structure. The G-quartzyme behaves as an allosteric ribozyme, with the potassium cations acting as positive effectors with a Hill coefficient of 2.9 ± 0.2. The conformation transition caused by the presence of the potassium ions is supported by enzymatic and chemical probing of both the inactive (*off*) and active (*on*) structures. This work shows that it is possible to interfere with the tight structure of the hepatitis delta virus ribozyme by adding an unusual, stable structure. To our knowledge, the G-quartzyme is the sole ribozyme that exhibits a monovalent cation dependent activity.

Keywords: Allosteric ribozyme; G-quadruplex; Catalytic RNA; Molecular engineering

INTRODUCTION

Molecular engineering of the hepatitis delta virus (HDV) ribozyme led to the creation of an allosteric ribozyme (Rz) possessing an endoribonuclease activity that is under the control of a G-quardruplex structure. HDV RNA strand that includes a self-cleaving RNA motif that has been separated into two molecules in order to develop *trans*-acting systems in which one molecule, identified as the ribozyme (HDV Rz), possesses the catalytic properties required to successively cleave several molecules of substrate (S) (Shih and Been, 2002). HDV Rz folds into a tertiary structure that is extremely stable and that retains its activity at temperatures as high as 80°C, as well as in denaturing buffer containing either 5 M urea or 18 M formamide (Doherty and Doudna, 2000; Shih and Been, 2002). Unlike other ribozymes, HDV Rz does not appear able to follow alternative folding pathways, an observation that, at least in part, could be due to the comparatively limited flexibility of its tightly packed structure (Doherty and Doudna, 2000; Lévesque et al. 2002; Krasovska et al. 2007). We wondered if it might be possible to alter the activity of a model, antigenomic, HDV Rz for which both the kinetic and thermodynamic behaviours have been extensively studied (Mercure et al. 1998; Ouellet and Perreault, 2004).

It is well documented that guanine-rich nucleic acid sequences can adopt a four-stranded helical structure termed a G-quadruplex, *in vitro* as well as under physiologically ionic conditions (Sen and Gilbert, 1988; Keniry, 2001; Davis, 2004; Saccà et al. 2005; Burge et al. 2006). The primary building block of this structure, called a G-quartet, is composed of four co-planar guanines that form Hoogsteen base pairs involving a total of eight hydrogen bonds (Gellert et al., 1962). These blocks then stack one on top of another forming a very stable helical G-quadruplex structure. Because each guanine positions its O_6 carbonyl group in the center of the G-quartet, there is an absolute requirement for the presence of a counter ion,

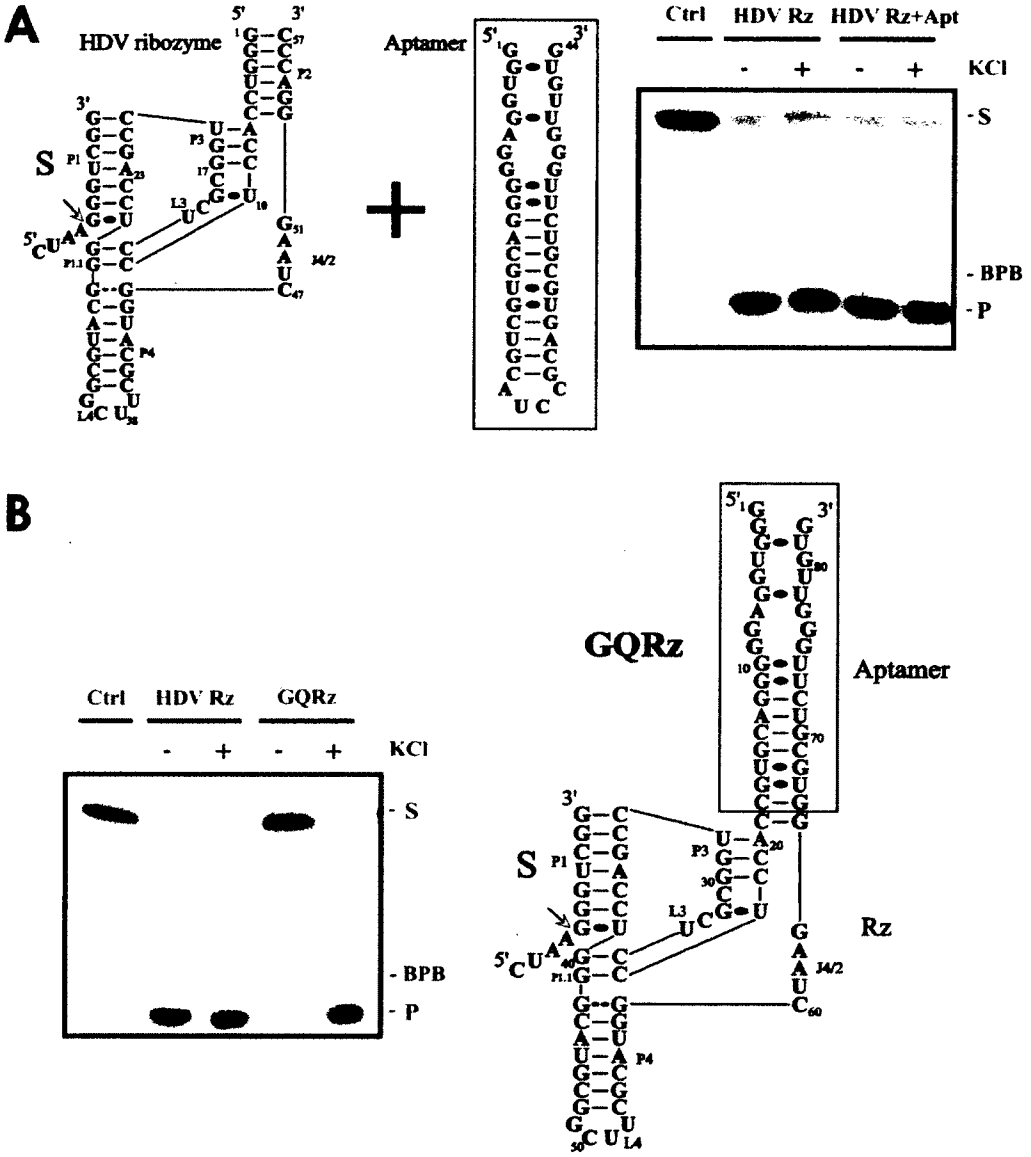
which typically is potassium (Hardin et al. 1997; Wong and Wu, 2003; Neidle and Balasubramanian, 2006).

Here, we reported the molecular engineering of a chimeric G-quadruplex – HDV ribozyme. This produced an allosteric ribozyme that is under the control of the potassium, an entirely new class of catalytic RNA. Structural and kinetic characterizations were performed in order to decipher the mechanism that activates this novel ribozyme.

RESULTS AND DISCUSSIONS

Engineering a G-quadruplex-HDV ribozyme

Initially, we decided to use an *trans*-acting version of the HDV ribozyme for which the kinetic behavior has been characterized under both single- and multiple-turnover conditions (e.g. see Mercure et al. 1998; Ananvoranich et al. 1999). This antigenomic HDV Rz efficiently cleaves a 5'-end ^{32}P-labelled 11 nucleotide substrate in the presence of 10 mM $MgCl_2$ in 1 hr at 37°C regardless of the presence or the absence of either 150 mM KCl, or of an independent RNA aptamer. In the absence of KCl, this aptamer was shown to fold into a rod-like formation of two RNA strands that include several consecutive guanosines joined by a hairpin (Figure 1A; Lévesque et al., 2007). In the presence of K^+ this aptamer was demonstrated to fold into a G-quadruplex structure whose precise structure remains to be solved. When this aptamer was then inserted in the P2 stem of the HDV ribozyme, a stem that is located outside of the catalytic core, cleavage activity was observed only in the presence of KCl (Figure 1B, lanes 4 and 5). In the absence of KCl, only trace amounts of product were detected, suggesting that the G-quadruplex has to be correctly folded in order for the ribozyme to exhibit any cleavage activity. Thus, this newly engineered ribozyme is now a G-quadruplex dependent one (namely G-quartzyme, GQRz), that is to say it represents a new class of Rz.


A
HDV ribozyme
Aptamer
Ctrl
HDV Rz
HDV Rz+Apt
KCl
S
BPB
P
B
GQRz
Aptamer
Rz
Ctrl
HDV Rz
GQRz
KCl
S
BPB
P
C
GQRz
GQRz Δ3'
GQRz Δ5'
GQRz-Δ3'-3G/A
GQRz-3'-G/A
−: 2 ± 1%
+: 86 ± 2%
−: 5 ± 1%
+: 5 ± 1%
−: 89 ± 4%
+: 86 ± 1%
−: 76 ± 1%
+: 73 ± 3%
−: 7 ± 2%
+: 10 ± 1%

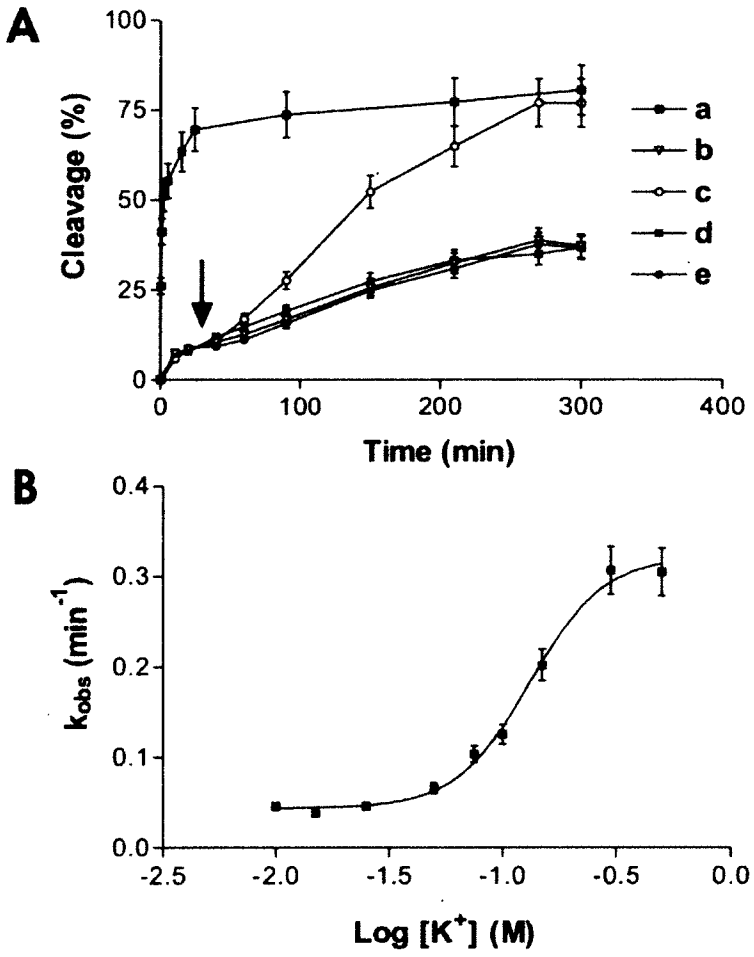
Chapitre 1, Figure 1. Characterization of the GQRz.

Secondary structure and nucleotide sequence of the HDV Rz and, in the box, the independent RNA aptamer (**A**), as well as of the GQRz (**B**). The aptamer is illustrated under its inactive structure as previously described (Lévesque et al. 2007). In both cases, autoradiograms of 20% polyacrylamide gels of cleavage assays performed with 10 nM Rz either in the presence (+) or absence (-) of 150 mM KCl are presented. The positions of the substrate (S), product (P) and bromophenol blue marker dye (BPB) are indicated on the right of the gel. The controls (Ctrl) are assays incubated in the absence of any ribozyme. (**C**) Sequence of the aptamer domain of the mutated GQRz and the percentages of cleavage activities observed after a 1 h incubation of 10 nM ribozyme in either the absence (-) or the presence (+) of 150 mM KCl. The boxed nucleotides denote the mutated ones.

In order to study the modulation of the ribozyme's activity caused by the introduction of the G-quadruplex structure, several mutants were synthesized (Figure 1C). Deletion of the 3'-strand of the aptamer domain led to a mutant that retained the low cleavage activity seen in the absence of KCl, but that could not be activated by the addition of KCl (GQRz-Δ3'). Conversely, deletion of the 5'-strand produced a mutant that is fully active regardless of the presence or absence of KCl (GQRz-Δ5'). Thus, this 5'-strand of the G-quadruplex seems to be involved in the inhibition. This conclusion receives physical support from the demonstration that the substitution of 3 guanosines for adenosines in GQRz-Δ3' led to the generation of a mutant that is active in the absence of KCl (GQRz-Δ3'-3G/A), showing that it is sequence specific. Finally, the mutation of a single guanosine in the 3'-strand to an adenosine led to an almost complete loss of the activation caused by the addition of KCl (GQRz-3'-G/A), illustrating the importance of the guanosines of the 3'-strand in the G-quadruplex formation.

Kinetic characterization of the GQRz

The kinetic behaviour of the GQRz was studied under various single turnover conditions. In the absence of KCl, it exhibited only residual cleavage in the presence of 50 mM $MgCl_2$ (<10%) while the original HDV Rz's cleavage level



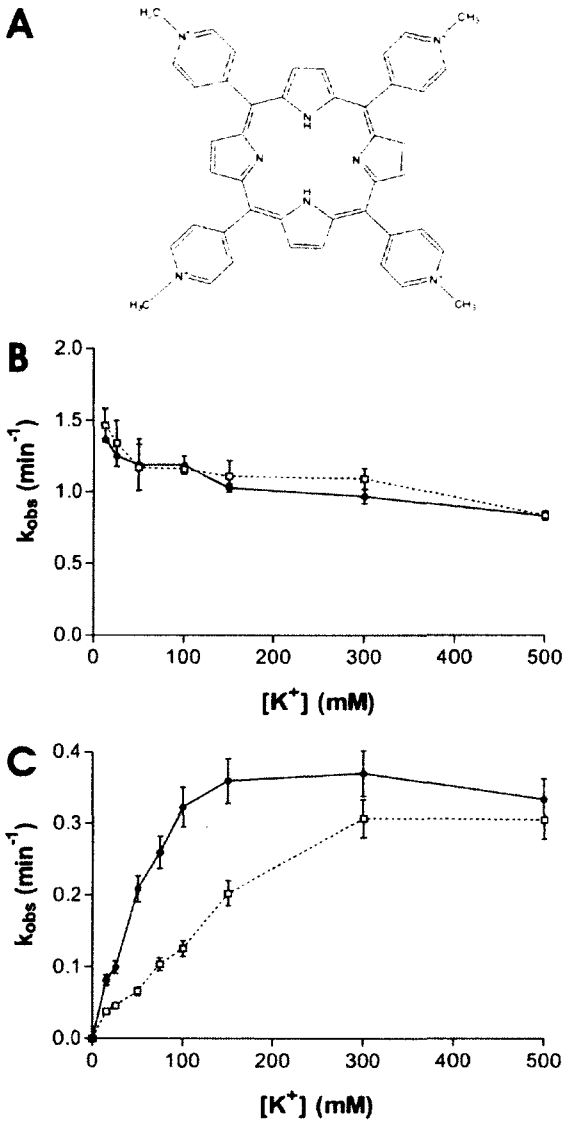
Chapitre 1, Figure 2. Kinetics characterization of the GQRz.

(**A**) Kinetic assays performed under single turnover conditions ([Rz 500 nM] > [S <1 nM]) with the HDV Rz (a) and the GQRz after the addition of water (b), 150 mM KCl (c), 150 mM NaCl (d) or 150 mM LiCl (e) after 30 min of incubation time (arrow mark). (**B**) Hill representation of cleavage assays using GQRz performed in the presence of 10 to 500 mM of KCl.

reached up to 75% after 30 min (Figure 2A). The addition of K^+ at this point caused a significant increase in the cleavage activity, while the addition of either Li^+ or Na^+, two monovalent ions that do not support formation of this G-quadruplex structure (Lévesque et al. 2007), did not. The addition of KCl led to cleavage levels equivalent to those of the original HDV Rz. The increased rate was slower than that of the original Rz because, in this assay, there was no slow-cooling step that

favoured rapid G-quadruplex formation. Kinetic constants were determined using various concentrations of Rz (5 to 1 600 nM) and trace amounts of substrate that were initially slow-cooled in the presence of 150 mM KCl in order to prefold the G-quardruplex. GQRz had k_{max} and $K_{M'}$ values of 0.77 ± 0.02 min^{-1} and 42.6 ± 7.1 nM, respectively, compared to 1.24 ± 0.01 min^{-1} and 13.6 ± 1.2 nM, respectively, for the original HDV Rz. The difference is likely primarily due being to the conformational transition. Both ribozymes had relatively similar magnesium dependences in the presence of KCl, with K_{Mg} of 23 ± 5 mM and 32 ± 6 mM for the GQRz and original Rz, respectively. The KCl dependence of the GQRz exhibited a sigmoid relationship characterized by a saturation at 150 mM and a Hill coefficient of 2.9 ± 0.2 (Figure 2B). Thus, this is an allosteric ribozyme whose activity is modulated by K^+ cations.

It has been well documented that G-quadruplexes are stabilized in the presence of porphyrin. Specifically, it has been demonstrated that the porphyrin generally stacks at the end of the G-quadruplex (Phan et al. 2005; Patel et al. 2007). Moreover, it has been shown that the fusion of a G-quadruplex to a hammerhead structure permitted modulation of the ribozyme cleavage activity by the porphyrin (Wieland and Hartig, 2006). The porphyrin was shown to be a novel RNA binder that exhibits the important inhibitory effect of an unmodified hammerhead ribozyme. However, the fusion of a G-quadruplex module to the hammerhead motif led to a ribozyme that turned the strongly inhibitory effect of the porphyrin into an activating one. In accordance with a previous report, it seems that the addition of the G-rich sequences creates a further binding site for the porphyrin. In order to support the idea that GQRz folds into a structure that includes G-quadruplex motifs, cleavage assays were performed in either the presence or the absence of 250 nM $TMPyP_4$ (meso-5,10,15,20-tetrakis-(N-methyl-4-pyridyl)porphine (a cationic porphyrin, Figure 3A) and various concentrations of KCl. In the case of the original ribozyme, the k_{obs} were virtually identical whether in the absence or the presence of $TMPyP_4$ (Figure 3B). Only small decreases of



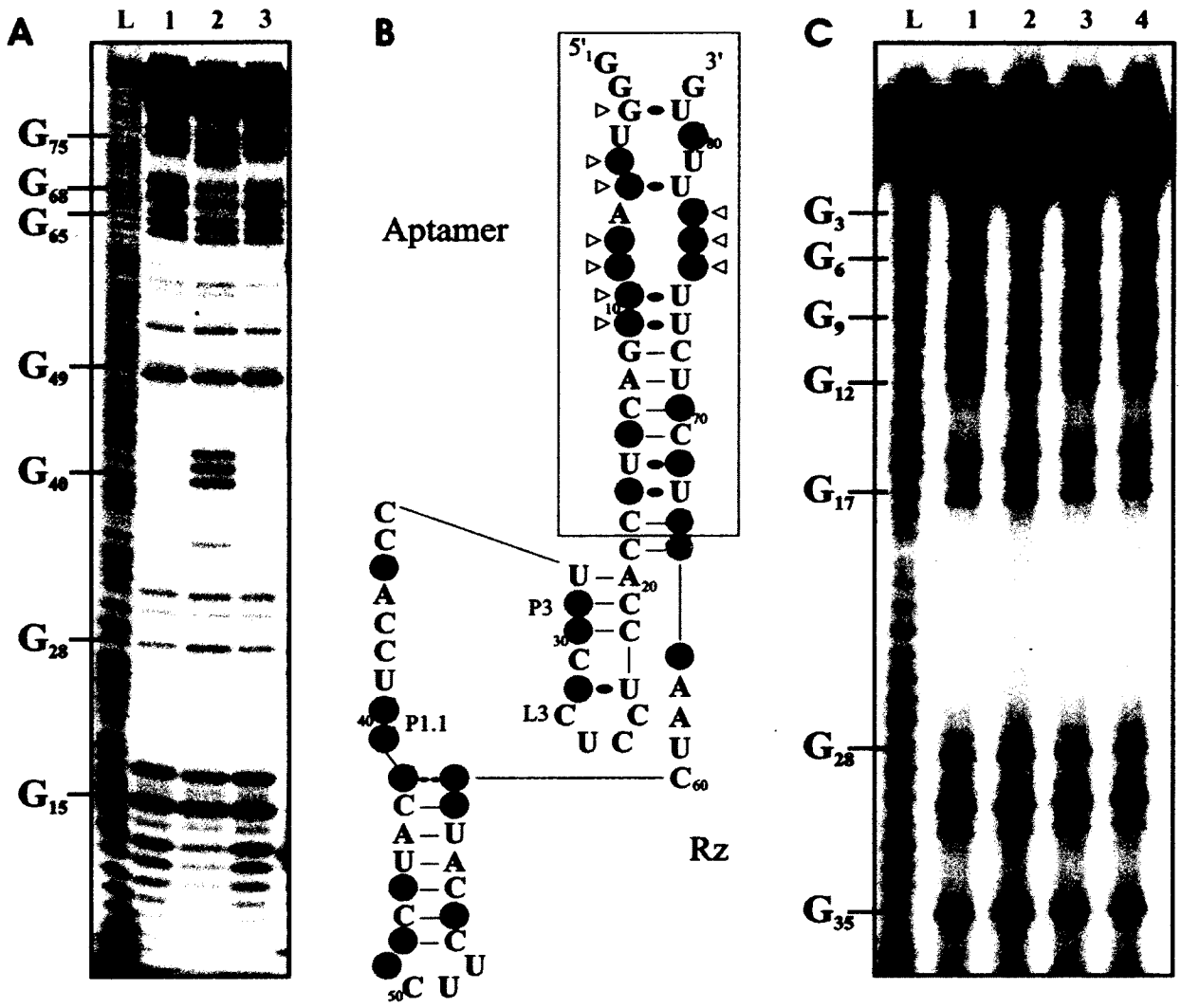
Chapitre 1, Figure 3. Impact of porhpyrin on the GQRz's activity.

Cleavage assays performed either in the absence or presence of porphyrin. (**A**) Molecular structure of the *meso*-5,10,15,20-tetrakis-(N-methyl-4-pyridyl)porphine ($TMPyP_4$). (**B**) and (**C**) are rate constants (k_{obs}) of the cleavage activity of the original ribozyme and GQRz, respectively, in various concentrations of KCl and either the absence (opened squares) or the presence of 250 nM $TMPyP_4$ (closed circles).

the cleavage activity were observed with the increase of KCl, most likely resulting from the K^+ and Mg^{2+} competing to bind to the ribozyme. Clearly, the situation was different with the GQRz. Higher concentrations of KCl were required in the absence of porphyrin in order to obtain the same level of GQRz activity observed in the presence of $TMPyP_4$ (Figure 3C). Moreover, kinetic analysis revealed that the Hill constant was reduced from 2.9 to 1.9 in the presence of $TMPyP_4$. These results support the formation of the G-quadruplex in the *on* state of the GQRz. Several experiments were performed to establish whether the G-quadruplex was resulting from intra- or from intermolecular interactions (data not shown). For example, the active GQRz was incubated with increasing concentration of the original aptamer, or inactive GQRz (resulting from mutations in the catalytic core of the ribozyme domain), and kinetics of cleavage performed. All the data obtained with or without prior heat denaturation-renaturation support the idea that the G-quadruplex results from intramolecular interactions, therefore GQRz would act as a monomer.

Probing the conformational transition

Structural differences caused by the addition of the K^+ ions were revealed by probing the guanosines. A typical autoradiogram for the ribonuclease T1 probing (RNase T1, an enzyme that preferentially cleaves single-stranded guanosines) of 5'-end labelled GQRz is shown in figure 4A. The observed banding pattern is virtually identical to that obtained when the GQRz was incubated in the presence of 150 mM of either LiCl or NaCl. However, upon the addition of 150 mM KCl several differences were observed. For example, the phosphodiester bonds of the guanosines located in positions 75, 76, 77 and 80 were no longer hydrolyzed. Data from several experiments using either 5'- or 3'-end labelled GQRz were compiled (Figure 4B). Briefly, this analysis led to two conclusions: i. globally, the structure of the ribozyme domain is in agreement with the proposed secondary structure, regardless of the presence or the absence of the KCl (i.e. all of the single-stranded



Chapitre 1, Figure 4. Structural characterization of the GQRz.

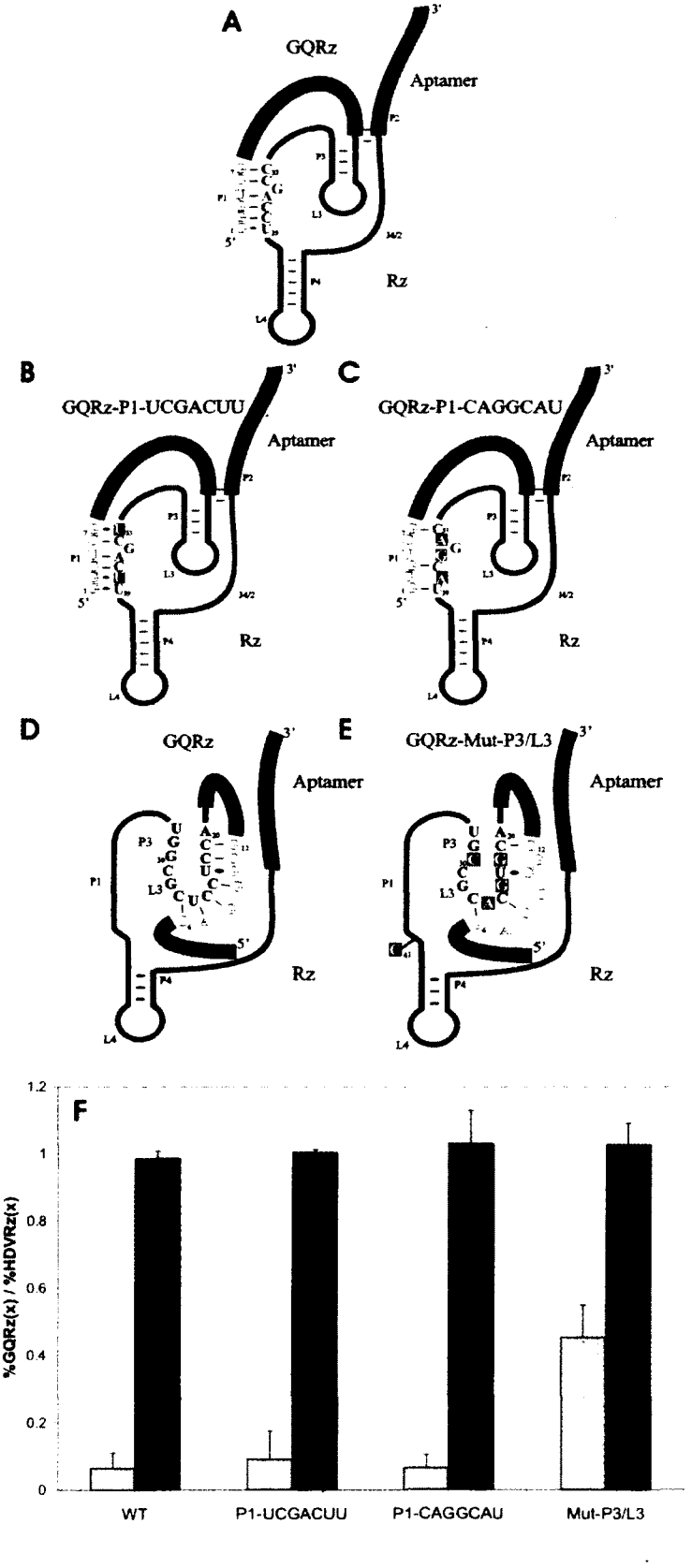
(**A**) Autoradiogram of a 8% denaturing (8 M urea) polyacrylamide gel of RNase T1 mapping of 5'-labelled GQRz performed in the presence of 150 mM of either NaCl (lane 1), KCl (lane 2) or LiCl (lane 3). Representative guanosine residues are indicated on the left of the gel. The lane designated L is an alkaline hydrolysis of GQRz. (**B**) Compilation of the mapping data. The circled guanosines were hydrolyzed by RNase T1 specifically in the absence (green) or presence (red) of KCl, or under both conditions (blue). The intensity of the color and the thickness of the characters are proportional to the relative amount of cleavage observed. The triangles indicate the guanosines protected from the DMS reagent in the presence of KCl. (**C**) Autoradiogram of an 8% denaturing (8 M urea) polyacrylamide gel of DMS probing performed with the 3'-end labelled GQRz in either the absence of monovalent salt (lane 1) or the presence of 150 mM of KCl (lane 2), NaCl (lane 3) or LiCl (lane 4).

guanosines were hydrolyzed, with the exception of those located in positions 40 to 42 that appear to be single-stranded only after the addition of KCl even though they should be single-stranded under both conditions); and, ii. the guanosines from the upper part of the aptamer domain were not hydrolyzed in the presence of KCl, supporting the notion that they are involved in the formation of the G-quadruplex. This hypothesis also received support from the observation that these guanosines were protected from DMS modification solely in the presence of KCl (see Figure 4B). The figure 4C shows a autoradiogram of a DMS probing performed with the 3'-end labelled GQRz. Specifically, this shows that the guanosine residues in positions 3, 5, 6, 8, 9, 10 and 11 were protected from the DMS modification solely in the presence of K^+.

Studies of the inactive conformation

The addition of KCl to GQRz has the effect of supporting a structural transition from an inactive (*off* state) to an active (*on* state) conformation of the ribozyme domain due to the formation of a G-quadruplex structure. It is likely that this transformation is possible because the formation of a G-quadruplex has been shown to favourably compete with the maintenance of existing Watson-Crick base pairs (Li et al. 2003).

The nature of the *off* state remains unclear; however, nuclease probing and sequence analysis suggested two possibilities that were subsequently verified by site-directed mutagenesis. Firstly, RNase T1 data revealed that the guanosines located in positions 40 to 42 that form the J1/4 junction that is single-stranded in the absence of substrate, a fact that is well supported by published data (e.g. see Ouellet and Perreault 2004), are not hydrolyzed in the off state, suggesting that they were either in a helical region or not accessible (Figure 4A). This might result from an interaction between nucleotides 1 to 7 of the aptamer domain and nucleotides 33 to 39 of the ribozyme domain as, with the exception of one residue,


A
GQRz
Aptamer
Rz
B
GQRz-P1-UCGACUU
Aptamer
Rz
C
GQRz-P1-CAGGCAU
Aptamer
Rz
D
GQRz
Aptamer
Rz
E
GQRz-Mut-P3/L3
Aptamer
Rz
F
%GQRz(x) / %HDVRz(x)
1.2
1
0.8
0.6
0.4
0.2
0
WT
P1-UCGACUU
P1-CAGGCAU
Mut-P3/L3

Chapitre 1, Figure 5. Characterization of the inactive conformation.

Cleavage assays performed with GQRz mutated in either the P1 stem or the P3-L3 stem-loop. (**A**) to (**C**) are schematic secondary structures showing the potential base-pairing between the 5'-strand of the aptamer domain and the P1 region of the ribozyme domain either in the original GQRz (**A**) or the two mutated versions (**B** and **C**). (**D**) and (**E**) are schematic secondary structures showing the potential base-pairing between the 5'-strand of the aptamer domain and the P3-L3 region of the ribozyme domain. In all cases, the mutated nucleotides are boxed and the aptamer is always represented under its inactive conformation. (**F**) Relative cleavage activities for the three versions after a 1 h incubation at 37°C under single single turnover conditions (100 nM Rz and 1 nM S) either in the presence (black bars) or the absence (light grey bars) of 150 mM KCl.

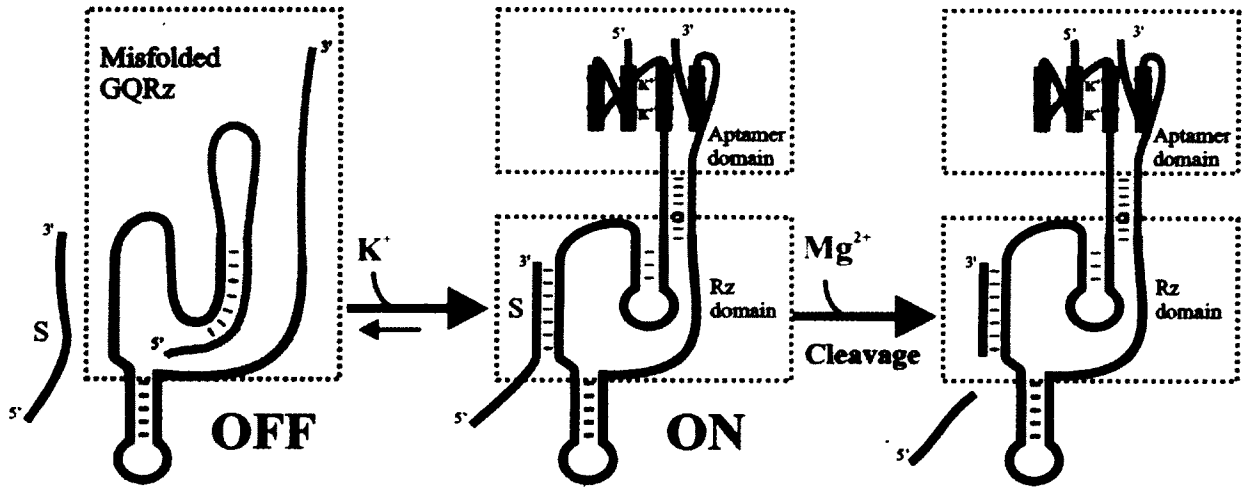
they are complementary. Two ribozymes mutated in the P1 stem, and the appropriate complementary substrates, were synthesized and their cleavage activities accessed (Figure 5, panels A-C and F). The mutants that could not support the formation of the potential interaction exhibited cleavage activities similar to that of the original sequence, thereby invalidating this hypothesis (Figure 5F). The panel F illustrates the results of the relative end-point determined for the GQRz mutant as compared to the original HDV ribozyme having the same mutations. Analysis of the k_{obs} values led to the same conclusion. Finally, a binding shift assay showed that the substrate bound to the GQRz in the absence of KCl, consequently, suggesting that the P1 region is not involved in the inhibitory mechanism (data not shown).

Secondly, nucleotides 6 to 12 of the aptamer can base-pair with nucleotides 21 to 27 of the P3-L3 stem-loop (Figure 5D). This binding is supported by RNase A probing experiments (data not shown), which preferentially hydrolyzes single-stranded C and U residues, suggesting that positions C21 to U26 were single-stranded in the *on* state, but not in the *off* state. Based on results from *in vitro* selection showing potential sequence variations in this region (Nehdi et al. 2007), a mutant GQRz was produced and tested (Figure 5E-F). The resulting GQRz-mutP3L3 showed a significant increase in its cleavage activity in the absence of KCl as compared to that of the original ribozyme, supporting the hypothesis that

the *off* state was due to this second potential interaction. The formation of this interaction may also support a different folding of the P1 stem, which would be responsible for the non-hydrolysis of the J1/4 guanosines by the RNase T1 (e.g. $_{31}$GUCCG$_{35}$ with $_{39}$UGGGC$_{43}$).

CONCLUDING REMARK

Altogether, these results led us to propose a model for the GQRz modulation in which the addition of the effector KCl supports a structural transition from an inactive (*off* state) to an active (*on* state) conformation of the ribozyme domain that depends on the formation of a G-quadruplex (see Figure 6). The nature of the *off* state remains unclear; however, preliminary experiments suggest that it likely results from base-pairing interactions between the 5'-strand of the aptamer domain and the L3-P3 stem-loop of the ribozyme domain. This work showed that it was possible to interfere with the tightly structured HDV Rz by the addition of an unusual structure. To our knowledge, the GQRz is the sole ribozyme possessing a monovalent cation dependent activity.



Chapitre 1, Figure 6. Model of the molecular mechanism of the GQRz.

MATERIALS AND METHODS

RNA synthesis

All ribozymes were synthesized by *in vitro* transcription using T7 RNA polymerase as described previously (Nehdi et al. 2007). Briefly, two overlapping oligonucleotides (2 µM each) were annealed and double-stranded DNA was obtained by filling in the gaps using purified *Pfu* DNA polymerase. The double-stranded DNA was then ethanol precipitated. The resulting DNA templates contained the T7 RNA promoter sequence followed by the ribozyme sequence. After dissolution in ultrapure water, run-off transcriptions were carried out in a final volume of 100 µL containing purified T7 RNA polymerase (10 µg) in the presence of RNA Guard (24 U, Amersham Biosciences), pyrophosphatase (0.01 U, Roche Diagnostics) and 5 mM NTP in a buffer containing 80 mM HEPES-KOH, pH 7.5, 24 mM $MgCl_2$, 2 mM spermidine, 40 mM DTT and using the double-stranded DNA as template. After a 2 h incubation at 37°C, the mixtures were treated with DNase RQ1 (Promega) at 37°C for 20 min, and the RNA then purified by phenol:chloroform extraction followed by precipitation with ethanol. The transcripts were fractionated by denaturing (8 M urea) 8% polyacrylamide gel electrophoresis (PAGE; 19:1 ratio of acrylamide to bisacrylamide) using 45 mM Tris-borate, pH 7.5/1 mM EDTA solution as running buffer. The RNAs were visualized by UV shadowing. The bands corresponding to the correct sizes of the ribozymes were excised from the gel and the transcripts eluted overnight at room temperature in buffer containing 1 mM EDTA, 0.1% SDS and 0.5 M ammonium acetate. The ribozymes were then ethanol precipitated, dried and dissolved in water. The concentrations were determined by absorbance at 260 nm.

Chemically synthesized ribooligonucleotides (RNA substrates: 5'-CUAAGGGUCGG-3', 5'-CUAAGAGUCGA-3' and 5'-CUAAGUGCCUG-3') were purchased from Integrated DNA Technologies (IDT).

RNA labelling

In order to produce 5'-end labelled ribozymes, transcripts were dephosphorylated by adding 1 U of antartic phosphatase (New England Biolab) to 50 pmol of ribozyme and incubating the reaction mixture for 30 min at 37°C in a final volume of 10 µL containing 50 mM Bis-Propane, pH 6.0, 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and 40 U of RNAGuard (Amersham Biosciences). The enzyme was then inactivated by incubation at 65°C for 5 min. Dephosphorylated ribozymes (10 pmol), or chemically synthesized RNA substrates, were 5'-end radiolabelled using 3 U of T4 polynucleotide kinase (Promega) at 37°C for 1 h in the presence of 3.2 pmol of [γ-^{32}P]ATP (6000 Ci/mmol, New England Nuclear). The reactions were stopped by the addition of ice-cold formamide dye buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol), and the RNA molecules purified by 8–20% polyacrylamide gel electrophoresis and recovered as described above except that the detection was performed by autoradiography.

For the 3'-end labelling of ribozymes, 15 pmol were incubated for 1 h at 37°C in a final volume of 10 µL containing 50 mM Tris-HCl, pH 7.8, 10 mM $MgCl_2$, 10 mM DTT, 1 mM ATP, 10% dimethylsulfoxide, 10 pmol [^{32}P]Cp (3000 Ci/mmol, New England Nuclear) and 10 U of T4 RNA ligase (New England Biolabs). After incubation, the ribozymes were purified on denaturing gels and recovered as described above.

Cleavage reactions and kinetics

Unless otherwise state, the cleavage reactions were carried out in 20 µL reaction mixtures containing 50 mM Tris-HCl, pH 7.5 either in the presence or the absence of 150 mM KCl and 10 mM $MgCl_2$ at 37°C for 1 h under single turnover conditions ([Rz] >> [S]). Prior to the reactions, trace amounts of 5'-^{32}P-end labeled substrates (<1 nM) and non radioactive ribozymes (10 nM) were mixed together in the Tris-HCl buffer and KCl (depending on the experiment), heated at 70°C for 2 min, slow-cooled to room temperature and then incubated at 37°C for 5 min. The cleavage

reactions were initiated by the addition of $MgCl_2$. After an incubation of 1 h at 37°C, the reactions were quenched by the addition of ice-cold formamide dye buffer. The mixtures were fractionated on denaturing 20% PAGE gels and exposed to PhosphorImager screens (Molecular Dynamic). The extents of cleavage were determined from measurements of the radioactivity present in both the substrate and the 5' product bands using the ImageQuant software. For time-course experiments, aliquots of 2.3 µL were taken at various times up to 1 h (5 h for the experiments that were not slow-cooled experiments) and were treated as described above.

Kinetic analyses were performed under single turnover conditions in the presence of 50 mM $MgCl_2$ and 150 mM KCl. Trace amounts of 5'-^{32}P-end labelled substrate (<1 nM) were cleaved using various concentrations of ribozyme (5-1600 nM), $MgCl_2$ (1-100 mM) or KCl (10-500 mM). Cleavage assays with various concentrations of KCl were also performed in the presence of the 250 nM $TMPyP_4$ (Calbiochem). The fractions cleaved were determined, and the rate of cleavage (k_{obs}) obtained from fitting the data to the equation $A_t = A_0(1-e^{-kt})$ where At is the percentage of cleavage at time t, A0 is the maximum percent cleavage (or the end-point of cleavage), and k is the rate constant (k_{obs}). Each rate constant was calculated from at least two independent measurements. The values of k_{obs} obtained were then plotted as a function of either the ribozyme or the Mg^{2+} concentration for the determination of k_{max}, $K_{M'}$ (i.e. a pseudo Michealis Menten constant which is more appropriate than a KD because the kinetic mechanism of HDV ribozyme involves several conformational transitions) and K_{Mg}. Using the GraphPad Prism software, the Hill coefficient was calculated by curve fitting using a sigmoidal dose-response with variable slope: Y = Bottom + (Top-Bottom)/(1+10$^{((LogEC50-X)Hillslope)}$) where Y is the k_{obs} and X is the logarithm of the KCl concentration in molar. The values obtained from independent experiments varied by less than 15%.

Enzymatic and chemical probing

RNase T1 probing was carried out with trace amounts of the 5'-^{32}P–end labelled ribozyme (<1 nM) supplemented with 1.5 µM of non-radioactive ribozyme and dissolved in 10 µL of buffer containing 20 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$ and 150 mM of either NaCl, KCl or LiCl. The ribozymes were heat-denatured and slow-cooled prior the addition of the magnesium. The mixtures were incubated for 2 min at 37°C in the presence of 0.6 U of RNase T1 (Roche Diagnostic), and were then quenched by the addition of 10 µL of ice-cold formamide dye buffer. For alkaline hydrolysis, the ribozymes (<1 nM) were dissolved in 5 µL of water, 1 µL of 1 N NaOH was added and the reaction incubated at room temperature for 1 min prior to being quenched by the addition of 3 µL of 1 M Tris-HCl, pH 7.5. The RNA molecules were then ethanol precipitated and dissolved in loading buffer. All samples were analyzed on denaturing 8% PAGE gels and visualized by exposure to PhosphorImager screens.

For chemical probing, 1 µL of DMS (dimethyl sulfate; diluted 1:8 in 100% ethanol) was added to the sample and then incubated at room temperature for a further 20 min. The RNA samples were ethanol precipitated, and the pellets washed twice with ethanol in order to remove all traces of DMS. The resulting pellets were dissolved in 20 µL of 500 mM Tris/HCl (pH 7.5). Sodium borohydride (200 mM; 10 µL) was added to the samples, which were then kept on ice for 5 min in the dark. Next, 10 µL of aniline solution (aniline/glacial acetic acid/water, 10:6:93, by vol.) was added to the samples and the tubes incubated at 60°C for 10 min in the dark. The ribozyme was then ethanol precipitated, fractionated on denaturing PAGE and analyzed.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Canadian Institutes of Health Research (CIHR, MOP-44002) to J.P.P. The RNA group is supported by a grant from the CIHR and the Université de Sherbrooke.. J.D.B. was supported by a pre-doctoral fellowship from CIHR. J.P.P. holds the Canada Research Chair in Genomics and Catalytic RNA.

REFERENCES

Ananvoranich, S., Lafontaine, D.A. and Perreault, J.P. 1999. Mutational analysis of the antigenomic trans-acting delta ribozyme: The alteration of the middle nucleotides located on the P1 stem. *Nucleic Acids Res.* **27:** 1473-1480.

Burge, S., Parkinson, G. N., Hazel, P.,Todd, A. K. and Neidle, S. 2006. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34:** 5402-5415.

Davis, J. T. 2004. G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. *Angew. Chem. Int. Ed. Engl.* **43**: 668-698.

Doherty, E.A. and Doudna, J.A. 2000. Ribozyme structures and mechanisms. *Annu. Rev. Biochem.* **69:** 597-615.

Gellert, M., Lipssett. M.N. and Davies, D.R. 1962. Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. USA* **48***:* 2013-2018.

Hardin, C. C., Corregan, M. J., Lieberman, D. V. and Brown, B. A. 1997. Allosteric interactions between DNA strands and monovalent cations in DNA quadruplex assembly: thermodynamic evidence for three linked association pathways. *Biochemistry* **36:** 15428-15450.

Keniry, M.A. 2001. Quadruplex structures in nucleic acids. *Biopolymers* **56:** 123-146.

Krasovska, M.V., Sefcikova, J., Reblova, K., Schneider, B., Walter, N.G. and Sponer, J. 2006. Cations and hydration in catalytic RNA: molecular dynamics of the hepatitis delta virus ribozyme. *Biophys. J.* **91:** 626-638.

Lévesque, D., Beaudoin, J.D., Roy, S. and Perreault, J.P. 2007. In vitro selection and characterization of RNA aptamers binding thyroxine hormone. *Biochem. J.* **403:** 129-138.

Lévesque, D., Choufani, S. and Perreault, J.P. 2002. Delta ribozyme benefits from a good stability in vitro that becomes outstanding in vivo. *RNA* **8:** 464-477.

Li, W., Miyoshi, D., Nakano, S. and Sugimoto, N. 2003. Structural competition involving G-quadruplex DNA and its complement. *Biochemistry* **42:** 11736-11744.

Mercure, S., Lafontaine, D.A., Ananvoranich, S. and Perreault, J.P. 1998. Kinetic analysis of delta ribozyme cleavage. *Biochemistry* **37:** 16975-16982.

Nehdi, A., Perreault, J., Beaudoin, J.D. and Perreault, J.P. 2007. A novel structural rearrangement of hepatitis delta virus antigenomic ribozyme. *Nucleic Acids Res.* **35:** 6820-6831.

Neidle S. and Balasubramanian, S. 2006. Quadruplex nucleic acids. RSC Publishing (Biomolecular Sciences)Cambridge, UK.

Ouellet, J. and Perreault, J.P. 2004. Cross-linking experiments reveal the presence of novel structural features between a hepatitis delta virus ribozyme and its substrate. *RNA* **10:** 1059-1072.

Patel, D.J., Phan, A.T. and Kuryavyi, V. 2007. Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, epub ahead of printing.

Phan, A.T., Kuryavyi, V., Gaur, H.Y. and Patel, D.J. 2005. Small-moleculart interaction with a five-guanine-tract G-quadruplex xtrucutre from the human MYC promoter. *Nat. Chem. Biol.* **1:** 167-173.

Saccà, B., Lacroix, L. and Mergny, J.L. 2005. The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic Acids Res.* **33:** 1182-1192.

Sen, D. and Gilbert, W. 1998. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature* **334:** 364-366.

Shih, I.H. and Been, M.D. 2002. Catalytic strategies of the hepatitis delta virus ribozymes. *Annu. Rev. Biochem.* **71:** 887-917.

Wieland, M. and Hartig, J.S. 2006. Turning inhibitors into activators: a hammerhead ribozyme controlled by a guanine quadruplex. *Angew. Chem. Int. Ed.* **45:** 5875-5878.

Wong, A. and Wu, G. 2003. Selective binding of monovalent cations to the stacking G-quartet structure formed by guanosine 5'-monophosphate: a solid-state NMR study. *J. Am. Chem. Soc.* **125:** 13895-13905.

CHAPITRE 2: Les G-quadruplexes présents dans les 5'-UTR des ARNm du transcriptome humain.

ARTICLE: 5'-UTR G-quadruplex structures acting as translational repressors

Jean-Denis Beaudoin and Jean-Pierre Perreault

Article publié dans: *Nucleic Acids Research* (**2010**) 38(20): 7022-7036

AVANT-PROPOS:

J'ai conceptualisé le projet et réalisé 100% des expériences rapportées dans cet article. J'ai également monté toutes les figures et participé de façon très active à l'écriture du manuscrit.

RÉSUMÉ

Puisque plus de 90% du génome humain est exprimé, il est logique d'assumer que les mécanismes de régulation post-transcriptionnelle soient le moyen primaire contrôlant la quantité d'information provenant de l'ARNm vers la protéine. Cette étude décrit une approche solide incluant des expériences *in silico*, *in vitro* et *in cellulo* permettant une évaluation à grande échelle de l'impact des G-quadruplexes comme répresseurs traductionnelles. Des séquences incluant des G-quadruplexes potentiels ont été sélectionnées parmi neuf gènes distincts encodant pour des protéines impliquées dans divers processus biologiques. Leurs habilitées à former des structures G-quadruplexes *in vitro* ont été évaluées en utilisant le dichroïsme circulaire, la dénaturation thermique et, nouvellement, le "*in-line probing*". Six séquences ont été observées pour se replier en des structures G-quadruplexes *in vitro*. De plus, toutes ces dernières ont montré une inhibition de la traduction *in cellulo* lorsque liées à un gène rapporteur. Des analyses de séquence, de mutagénèse dirigée et d'autres expériences ont été effectuées dans le but de définir des règles régissant le repliement des G-quadruplexes. En outre, l'impact de polymorphisme nucléotidique simple (SNP) a été démontré pour être important dans la formation de G-quadruplexes localisés dans la région 5'-UTR de l'ARNm. À la lumière de ces résultats, les G-quadruplexes dans les 5'-UTR représentent clairement une classe de répresseurs traductionnels qui sont largement distribués dans la cellule.

ABSTRACT

Given that greater than 90% of the human genome is expressed, it is logical to assume that post-transcriptional regulatory mechanisms must be the primary means of controlling the flow of information from mRNA to protein. This report describes a robust approach that includes *in silico*, *in vitro* and *in cellulo* experiments permitting an in-depth evaluation of the global impact of G-quadruplexes as translational repressors. Sequences including potential G-quadruplexes were selected within 9 distinct genes encoding proteins involved in various biological processes. Their abilities to fold into G-quadruplex structures *in vitro* were evaluated using circular dichoism, thermal denaturation and the novel use of in-line probing. Six sequences were observed to fold into G-quadruplex structures *in vitro*, all of which exhibited translational inhibition *in cellulo* when linked to a reporter gene. Sequence analysis, direct mutagenesis and subsequent experiments were performed in order to define the rules governing the folding of G-quadruplexes. In addition, the impact of single nucleotide polymorphism was shown to be important in the formation of G-quadruplexes located within the 5'-untranslated region of an mRNA. In light of these results, clearly the 5'-UTR G-quadruplexes represent a class of translational repressors that is broadly distributed in the cell.

INTRODUCTION

The life cycle of a mRNA species is full of diverse processing events and regulatory controls. For a long time it has been believed that the primary means of regulating gene expression occurred at the transcription level. However, the discovery that over 90% of the genome is transcribed prompted the conclusion that post-transcriptional regulation is in fact the cornerstone for the regulation of gene expression [1]. Post-transcriptional regulatory elements must be involved in order to direct the expression of specific subsets of genes within this large transcriptome. In terms of the mRNAs themselves, these regulatory elements can act at various steps in their life cycles, ranging from their processing events (e.g. capping, splicing and polyadenalytion) to their active transport, stability and translation [2]. Several cellular factors are involved in these regulatory mechanisms. Some of them act as trans-acting regulatory elements. This is the case for the micro-RNAs, which generally interact with the 3'-untranslated regions (3'-UTR) of specific mRNAs, repressing their translation and/or decreasing their stabilities [3,4]. There are also many *cis*-acting regulatory factors. In general, the latter are highly ordered RNA structures present in either the 5'- or 3'-UTRs. For example, the presence of a highly active hammerhead ribozyme in the 3'-UTRs of the rodent C-type lectin type II gene has been shown to reduce protein expression in mouse cells [5]. Moreover, riboswitches which are implicated in regulating gene expression have been detected in the 5'-UTRs of a large variety of genes. Specifically, the binding of a metabolite to the aptamer domain has the effect of controlling the gene's expression level, leading to either an increase or a decrease in the transcription and/or the translation levels. New riboswitches are frequently discovered, and both their complexities and diversities remain unappreciated [6]. Clearly, the discovery and elucidation of post-transcriptional regulatory elements represent key components in achieving a good understanding of the molecular biology of the cell.

Guanine-rich nucleic acid sequences can fold into a non-canonical tetrahelical structure called a G-quadruplex. This structure involves the stacking of hydrogen-bonded G-tetrads, which, once stabilized by the chelation of monovalent metal ions such as potassium, represent an extremely stable four-stranded helical structure [7,8,9]. Several bioinformatic studies have revealed both a significant level of conservation and an enrichment in potential G-quadruplex (PG4) sequences in various regulatory elements (e.g. telomeres, DNA promoters and both the 5'- and 3'-UTRs of mRNAs) within the genome, suggesting that they are involved in key biological processes [10,11,12,13,14]. For example, the formation of G-quadruplexes at the eukaryotic telomeric sequences has been proposed to be associated with the telomeres' maintenance by modulating their interactions with various proteins [15,16,17]. The prevalence of PG4s within different functional classes of genes was determined using a computational approach [18]. For example, many G-quadruplexes have been found to be located in the promoters of various proto-oncogenes such as *c-MYC*, *C-Kit*, *c-myb* and *KRAS* [19,20,21,22]. These PG4 sequences were suggested to be involved in the regulation of the transcriptional activity of these genes. Moreover, because the G-quadruplexes are directly linked to several key features in cancer cells, such as telomeres and oncogenes, great efforts have been made to try and find potential ligands that would act as anticancer agents. Some compounds that were shown to target DNA G-quadruplexes have already provided promising results, either by inhibiting the telomerase activity, or by reducing oncogene expression [23].

While our knowledge of the DNA G-quadruplexes present in the human genome is increasing, our understanding of biologically relevant RNA G-quadruplexes remains limited. It is known that for a given sequence *in vitro*, an RNA G-quadruplex is usually more stable than its DNA counterpart [24]. Moreover, unlike DNA which is constrained mainly to a duplex form in the cell, RNA has no complementary strand limiting its structure. These two features make G-rich RNA sequences more susceptible to folding into a G-quadruplex structure *in vivo*. Several bioinformatic analyses searching for PG4 in the different regions of an

mRNA have been reported (e.g. in the 5'-UTR, the 3'-UTR and the RNA processing sites) [12,25]. Moreover, in some cases, RNA G-quadruplexes have been demonstrated to have functional roles [26,27,28,29,30,31,32]. For instance, one G-quadruplex structure was shown to direct the discrimination of a proper target by the fragile X mental retardation protein, while another was reported to regulate an alternative splicing event, to name two examples [26,27]. The original study showing a G-quadruplex structure acting as a translational repressor was performed in a cell free system using the full length *NRAS* 5'-UTR that includes such a structure [29]. Subsequently, two other studies showed similar effects *in cellulo* using either a 27 nucleotide *Zic-1* RNA G-quadruplex, or a complete *MT3-MPP* 5'-UTR bearing a special purine-only RNA G-quadruplex [30,31]. In each of these studies only one RNA G-quadruplex was analyzed. More recently, the characterization of artificial *cis*-acting G-quadruplex repressors revealed an interesting correlation between the loop length and the number of G-tracks in terms of the translational inhibition level [32]. Despite all of these studies, both the global impact and the importance of the 5'-UTR G-quadruplex structures on the biology of the cell remains, most likely, under estimated. Here, we present a robust approach including *in silico*, *in vitro* and *in cellulo* experiments that permits a wider evaluation of the G-quadruplexes acting as translational repressors. Importantly, several G-quadruplex structures widely distributed within the transcriptome were studied and new rules governing the formation of the G-quadruplexes are reported. These rules permit the proposal of several regulatory mechanisms of G-quadruplex formation in an RNA strand.

MATERIALS AND METHODS

The sequences of all of the oligonucleotides used in this work are given in Table S1.

Bioinformatics

The 5'-UTR databases were derived from sequences taken from Transterm and UTRdb [33,34]. These two databases contain spliced 5'-UTR sequences. PG4 sequences were identified using the above algorithm and the program RNAMotif [35]. The results were subjected to various homemade Perl scripts and manually cured in order to obtain the PG4 databases presented in the supplementary data in an Excel file format. When a 5'-UTR PG4 was identified in a gene that generates more than one transcript with the same 5'-UTR, each transcript was treated individually and was counted as one more PG4. The gene ontology analysis was performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID) web-accessible programs [36]. The input data for the web-accessible program was the list of genes that included a PG4 in the complementary strand obtained from the human UTRfull database (Dataset S6). The SNP analysis was performed using a database of the SNPs present in various human mRNAs corresponding to NCBI dbSNP build 129, and the PG4 database obtained from the human UTRef database (Dataset S3). The presence of SNPs inside each PG4 sequence was examined using several homemade Perl scripts that compare the positions and the lengths of the PG4s to the positions of the SNPs present in the mRNAs. The list of SNPs found within the PG4 sequences was manually cured, and is presented in the supplementary data (Dataset S4).

RNA synthesis

All PG4 versions used for the *in vitro* experiments were synthesized by *in vitro* transcription using T7 RNA polymerase as described previously [37]. Briefly, two overlapping oligonucleotides (2 µM each) were annealed, and double-stranded DNA was obtained by filling in the gaps using purified *Pfu* DNA polymerase in the presence of 5% DMSO. The double-stranded DNA was then ethanol-precipitated. The resulting DNA templates contained the T7 RNA promoter sequence followed by the PG4 sequence. After dissolution of the PCR product in ultrapure water, run-off transcriptions were performed in a final volume of 100 µL using purified T7 RNA

polymerase (10 μg) in the presence of RNase OUT (20 U, Invitrogen), pyrophosphatase (0.01 U, Roche Diagnostics) and 5 mM NTP in a buffer containing 80 mM HEPES-KOH, pH 7.5, 24 mM $MgCl_2$, 2 mM spermidine and 40 mM DTT. The reactions were incubated for 2 h at 37°C. Upon completion, the reaction mixtures were treated with DNase RQ1 (Promega) at 37°C for 20 min. The RNA was then purified by phenol:chloroform extraction followed by ethanol precipitation with ethanol. RNA products were fractionated by denaturing (8 M urea) 10% polyacrylamide gel electrophoresis (PAGE; 19:1 ratio of acrylamide to bisacrylamide) using 45 mM Tris-borate, pH 7.5/1 mM EDTA solution as running buffer. The RNAs were visualized by UV shadowing, and those corresponding to the correct sizes of the PG4s were excised from the gel and the transcripts eluted overnight at room temperature in buffer containing 1 mM EDTA, 0.1% SDS and 0.5 M ammonium acetate. The PG4s were then ethanol-precipitated, dried and dissolved in water. The concentrations were determined by spectrometry at 260 nm.

Circular dichroism spectroscopy

All circular dichroism (CD) experiments were performed using 4 μM of the relevant RNA sample dissolved in 50 mM Tris-HCl (pH 7.5) either in the absence of monovalent salt, or in the presence of 100 mM LiCl, NaCl or KCl. Prior to taking the CD measurement, each sample was heated to 70°C for 5 min and then slow cooled to room temperature over a 1 h period. CD spectroscopy experiments were performed with a Jasco J-810 spectropolarimeter equipped with a Jasco Peltier temperature controller in a 1 mL quartz cell with a pathlength of 1 mm. CD scans, ranging from 220 to 320 nm, were recorded at 25°C at 50 nm min^{-1} with a 2 sec response time, 0.1 nm pitch and 1 nm bandwidth. The means of at least three wavelength scans were compiled. Subtraction of the buffer was not required since control experiments in the absence of RNA showed negligible curves. CD melting curves were obtained by heating the samples from 25°C to 90°C at a controlled

rate of 1°C min^{-1} and monitoring a 264 nm CD peak every 0.2 min. T_m values were calculated using "fraction folded" (θ) *versus* temperature plots [38].

RNA labeling

In order to produce 5'-end-labeled PG4s, purified transcripts were dephosphorylated by adding 1 U of antartic phosphatase (New England BioLabs,) to 50 pmol of RNA and incubating the reaction mixture for 30 min at 37°C in a final volume of 10 µL containing 50 mM Bis-Propane (pH 6.0), 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and RNase OUT (20 U, Invitrogen). The enzyme was inactivated by incubation for 5 min at 65°C. Dephosphorylated transcripts (5 pmol) were 5'-end-radiolabeled using 3 U of T4 polynucleotide kinase (Promega) for 1 h at 37°C in the presence of 3.2 pmol of [α-^{32}P]ATP (6000 Ci/mmol; New England Nuclear). The reactions were stopped by adding formamide dye buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025% xylene cyanol), and the RNA molecules purified by 10% polyacrylamide gel electrophoresis. The bands of the correct sizes containing the 5'-end-labeled RNAs were excised and recovered as described above except that the detection was performed by autoradiography.

In-line probing

5'-end-labelled RNA (50 000 cpm), that is to say a trace amount of RNA (<1 nM), was heated at 70°C for 5 min and then slow cooled to room temperature over 1 h in buffer containing 50 mM Tris-HCl (pH 7.5) and either no monovalent salt, or in the presence of 100 mM LiCl, NaCl or KCl in a final volume of 10 µL. Following this incubation, the final volume of each sample was adjusted to 100 µL such that the final concentrations were 50 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$ and either no salt or 100 mM LiCl, NaCl or KCl. The reactions were then incubated for 40 h at room temperature, ethanol-precipitated and the RNAs dissolved in ice cold formamide dye loading buffer (95 % formamide and 10 mM EDTA). For alkaline hydrolysis, 50 000 cpm of 5'-end-labeled RNA (<1 nM) were dissolved in 5 µL of water, 1 µL of 1 N NaOH added and the reactions incubated for 1 min at room temperature prior to

being quenched by the addition of 3 µL of 1 M Tris-HCl (pH 7.5). The RNA molecules were then ethanol-precipitated and dissolved in formamide dye loading buffer. An RNase T1 ladder was prepared using 50 000 cpm of 5'-end-labeled RNA (<1 nM) dissolved in 10 µL of buffer containing 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$ and 100 mM LiCl. The mixture were incubated for 2 min at 37°C in the presence of 0.6 U of RNase T1 (Roche Diagnostic), and was then quenched by the addition of 20 µL of formamide dye loading buffer. The radioactivity of the in-line probing samples and both ladders was calculated, and equal amounts in terms of cpm of all conditions and ladders of each candidate were fractionated on denaturing (8 M urea) 10% polyacrylamide gels.

Plasmid construction

The sequences of the 5'-UTRs were obtained from the NCBI database and correspond to the following Gene Identification (GI) for each candidate: *EBAG9* (GI: 37694064), *FZD2* (GI: 5922012), *BARHL1* (GI: 31542183), *NCAM2* (GI: 33519480), *THRA* (GI: 46255056), *AASDHPPT* (GI: 20357567) and *TNFSF12* (GI: 23510442). The full length 5'-UTRs of each candidate was reconstituted *in vitro* by the filling in of multiple overlapping oligonucleotides and various PCR steps (the specific sets of oligonucleotides used for each candidate are shown in Table S1). Wild type and G/A-mutant 5'-UTR versions were synthesized for each candidate. In addition to the G/A-mutants, both C/A-mutants and CG/AA-mutants were synthesized for *TNFSF12*, and a C7 SNP 5'-UTR version were synthesized for *AASDHPPT*. The positions of all of the different mutations are the same as those used for the *in vitro* experiments. The list of oligonucleotides used for each candidate is shown in Table S1. The reconstituted 5'-UTRs were inserted in the Nhe I site in the pRL-TK plasmid vector (Promega). DNA sequencing of each candidate confirmed the insertion of the correct sequence.

Cell culture

HEK 293 cells (human embryonic kidney) were cultured in T-75 flasks (Sarstedt) in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% FBS, 1 mM sodium pyruvate and an antibiotic-antimycotic drug mixture (all purchased from Wisent) at 37°C in a 5% CO_2 atmosphere in a humidified incubator.

Dual luciferase and quantitative RT-PCR assays

HEK 293 cells ($1.2x10^5$) were seeded in 24-well plates. Twenty-four hours later, the cells were co-transfected with both the specific pRL-TK plasmid construction (renillia luciferase, *Rluc*) and the pGL3-control vector (firefly luciferase, *Fluc*) (Promega) using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. Twenty-four hours after transfection, 10% of the cells were used to measure the Rluc and Fluc activities using the Dual-luciferase Reporter Assay kit (Promega) according to the manufacturer's protocol in a 5 ml test tube using a Berthold Lumat LB9501 luminometer (Berthold Technologies). For each lysate, the value of the Rluc was divided by the value of the Fluc. The ratios obtained for the G/A-mutant version were compared to those obtained with the wild type version of each candidate. Both the mean value and the standard deviation were calculated from at least three independent experiments for each candidate.

Total cellular RNA was extracted from the remaining cells using an Absolute RNA Microprep Kit (Stratagene) according to the manufacturer's protocol that include a DNase treatment. Total RNA (200 ng) from each sample was reverse transcribed using Transcriptor Reverse Transcriptase (Roche). The cDNA was subjected to quantitative real-time PCR using the FastStart Universal SYBR Green Master (Rox) mix (Roche) and a Rotor-Gene™ 3000 device (Corbett Research). The levels of *Rluc* and *Fluc* mRNAs were detected using the appropriate primers sets: forward primers *Rluc* 5'-(TGGGGTGCTTGTTTGGCATT)-3' and *Fluc* 5'-(AAATGTCCGTTCGGTTGGCA)-3' and reverse primers *Rluc* 5'-(TGGCAACATGGTTTCCACGA)-3' and *Fluc* 5'-

(ACTCCGATAAATAACGCGCCCA)-3'. The relative gene expression data was calculated using the $\Delta\Delta C_T$ with the Fluc gene as internal control and the wild type version as calibrator for each candidate [39].

RESULTS

Frequence of G-quadruplexes within 5'-UTR

In order to better understand the general role that G-quadruplexes play as translational repressors, a database of all potential G-quadruplexes located in the 5'-UTRs of the genes from 18 organisms, including humans, was constructed. The human 5'-UTR sequences were downloaded from both UTRdb and Transterm, while those from the other organisms (listed in Dataset S1) were downloaded only from Transterm [33,34]. Potential G-quadruplex (PG4) sequences were identified using a previously available algorithm that searches for the sequence G_x-N_{1-7}-G_x-N_{1-7}-G_x-N_{1-7}-G_x, where $x \geq 3$ and N is any nucleotide (A,C,G or U) [40,41]. These parameters were established by taking into account various results from *in vitro* studies on the G-quadruplex structure. With these guidelines, the 5'-UTR sequences were scanned in order to identify PG4s located on either the template or the complementary strand. The PG4s located on the template strands are composed of tracks of cytosines in the sequence database, while those located on the complementary strands correspond to tracks of guanosines and will be found in the mRNA. The primary analysis was focused on the 124 315 5'-UTRs obtained from the human UTRfull collection (Table 1 and Dataset S2). This yielded 9 979 (8.0%) 5'-UTRs that contained at least one PG4 sequence. The numbers of 5'-UTRs with PG4s located in the template, *versus* those located in the complementary strand, was slightly different (6 092 (4.9%) *versus* 5 027 (4.0%) sequences, respectively). In total, 17 844 PG4s were found in the 5'-UTR, and are unequally distributed between the two strands. A significantly smaller number of potential G-quadruplex structures was observed in the complementary strand, that is to say in the mRNA, as compared to the template DNA strand (40.3% *versus*

59.7% respectively), suggesting potential biological consequences. The same unequal strand distribution is observed in the 4 other species with greater than 100 PG4s identified, supporting this statement (see Dataset S1). Moreover, a previous study reported the same bias for the distribution of the PG4s between the template and complementary strands [12]. Another interesting observation was the higher PG4/5'-UTR ratio observed for the template strand, suggesting that the cell is better able to deal with consecutive G-quadruplexes in the template strand than in the mRNA (Table 1). However, some 5'-UTRs can contain up to 5 different PG4s in the complementary strand (e.g. *ANKRD30B*, *CAV2* and *CDKN2D*; Dataset S2). The PG4 density in 5'-UTRs was estimated to be 0.292/kbase for the template strand, and 0.198/kbase for the complementary strand. In both cases, it represents a significant enrichment (4- to 5-fold) as compared to the reported PG4 density of the human genome (0.057/kbase) using the same algorithm [12].

Chapitre 2, Table 1. Incidence of potential G-quadruplexes in a human 5'-UTR database.

	Template strand	Complementary strand	Total
Nb. of 5'UTR	-	-	124 315
Nb of 5'UTR with PG4 (%)	6 092 (4.9%)	5 024 (4.0%)	9 979 (8.0%)
5'UTR with 1 PG4 (%)	3 313 (54.4%)	3 399 (67.7%)	5 133 (51.4%)
5'UTR with more than 1 PG4 (%)	2 779 (45.6%)	1 625 (32.3%)	4 846 (48.6%)
Nb. PG4	10 646	7 198	17 844
% of PG4	59.7%	40.3%	-
Ratio PG4/5'UTR	1.75	1.43	1.79
PG4 density	0.292/kbase	0.198/kbase	0.490/kbase

Ability of the selected candidates to fold into G-quadruplex structures *in vitro*

With the goal of evaluating the global impact of the G-quadruplex structure on the transcriptome, several candidates were selected with which to continue this study. Specifically, 9 5'-UTRs bearing a PG4 on their complementary strand were chosen based on the bioinformatic analysis (Table 2). The main criterion of selection was that these candidates' mRNAs had to encode proteins important for various cellular pathways; therefore they constituted a good representation of gene heterogeneity. The first step was the demonstration of whether or not the candidate's PG4 sequences adopted a G-quadruplex structure *in vitro*. Three different biochemical methods were used with each candidate, providing a reliable evaluation of the situation. The experiments were performed using transcripts that exceed the PG4 sequence requirement (i.e. they are longer) in order to better reflect the biological context of each 5'-UTR instead of only considering the guanosine tracks. However, it was not possible to use the complete 5'-UTR sequence, due to both technical constraints and difficulties in analyzing the results. Consequently, in all *in vitro* experiments the sequence of a PG4 was flanked by approximately ~15 nucleotides both upstream and downstream, and began with at least two consecutive guanosines as these are required for efficient *in vitro* transcription (see Figure 1A for the detailed sequences). Moreover, a G/A-mutant created by mutating several guanosines into adenosines in such a way that it prevents the formation of a G-quadruplex structure was also synthesized for each candidate (Figure 1A). These mutants were used as negative controls.

Chapitre 2, Table 2. Gene ontology of the 9 candidate genes.

Gene	Function	Process
EBAG9	Apoptotic protease activator activity	Apoptosis/Regulation of cell growth
FZD2	G-protein coupled receptor activity	G-protein coupled receptor protein signaling pathway
BARHL1	Protein binding/Transcription factor activity	Midbrain development/Neuron migration
NCAM2	Protein binding	Cell adhesion/Neuron adhesion
THRA	Thyroid hormone receptor activity	Hormone-mediated signaling
AASDHPPT	Transferase activity	Macromolecule biosynthetic process
TNFSF12	Cytokine activity	Apoptosis/Cell differentiation
MAP3K11	JUN kinase kinase kinase activity	Regulation of JNK cascade/Cell proliferation
DOC2B	Calcium ion binding/Transporter activity	Transport

The first method used for detecting G-quadruplex formation was analysis by circular dichroism (CD). This is a classical technique that detects G-quadruplex structures possessing the typical spectrum caused by the topology of the four-stranded helical structure (i.e. parallel or anti-parallel). Due to the nature of its sugar, an RNA G-quadruplex structure is compelled to adopt a parallel form. More specifically, the ribose residues prefer the puckering $C_{3'}$-*endo* conformation. This in turn favors that the glycosidic bond of every guanosine involved in the core of G-tetrads be in the *anti* orientation [42]. The formation of a parallel G-quadruplex structure provokes the appearance of a negative peak at 240 nm and a positive one at 264 nm [43]. It is important to focus on the transition of both characteristic peaks, when comparing the spectra recorded under two different conditions, in order to propose that the RNA molecule forms a G-quadruplex. The analysis cannot rely on a single spectrum, because other RNA structural features exist that can give a positive peak around 260 nm, as this would lead to a potential false positive G-quadruplex signature. The CD spectra for each candidate were initially recorded either in the absence of salt, or in the presence of 100 mM LiCl, two conditions that do not support the formation of G-quadruplex structures. The presence of Li^+ is the most reliable control in order to identify the "intrinsic" or initial structure of the RNA molecule because it provides the same ionic force as Na^+ or K^+, but it cannot support the formation of a G-quadruplex structure. Subsequently,

the experiments were repeated in the presence of 100 mM of either NaCl or KCl, two conditions that should favor the formation of G-quadruplex structures. Panel B of Figure 1 shows the recorded CD spectra for the *EBAG9*-derived transcripts as an example of a result typical of one positive for G-quadruplex formation, while the panel 1C illustrates the corresponding G/A-mutant. Clearly, there is a significant transition to a higher positive peak at 264 nm, and a negative one at 240 nm, when using the wild type version in the presence of KCl. No corresponding transition was observed for the G/A-mutant. Six out of nine candidates exhibited CD spectra with G-quadruplex signatures. Specifically, the *BARHL1* and *NCAM2* PG4 sequences appear to fold into G-quadruplex structures in the presence of either KCl or NaCl, while the *EBAG9*, *FZD2*, *THRA* and *AASDHPPT* sequences adopt this structure solely in the presence of KCl. Conversely, the *TNFSF12*, *MAP3K11* and *DOC2B* PG4 sequences did not show any evidence of a G-quatruplex signature, regardless of the nature of the salt present in the buffer. Similarly, the G/A-mutants never exhibited a significant transition characteristic of the formation of a G-quadruplex structure.

With the goal of confirming that some of the PG4 sequences do indeed fold into G-quadruplexes, thermal denaturation studies were then performed. The formation of a G-quadruplex in the presence of an appropriate cation (e.g. Na^+ or K^+) should lead to an increased stability (i.e. a higher melting temperature, T_m) of the RNA molecule as compared to one containing a structure involving only Watson-Crick base pairs [44]. The six PG4 sequences that, according to CD analysis, folded into G-quadruplex structures were examined in this way. The presence of LiCl provokes only a small increase in the T_m as compared to the value obtained in the absence of salt (Table 3). This stabilization of the RNA structure is due to the counterion effect of the cations that reduces the repulsion of the negative charge of the phosphate backbone. Conversely, the presence of KCl in the solution led to a significant increase in the T_m of each PG4 sequence (Table 3). In fact, all of the RNA structures were incompletely denatured, even at 90°C. Clearly, this experiment provides additional physical evidence supporting the

conclusion that these 6 PG4 sequences adopt a G-quadruplex structure at a physiological KCl concentration (i.e. 100 mM). Finally, an increase in the T_m value was also observed in the presence of NaCl, but only for the *BARHL1* and *NCAM2* PG4-derived sequences, in agreement with the CD analysis. The G/A-mutants exhibited no increase in their T_m values under all conditions, indicating that their structures were not altered by the addition of a monovalent cation such as Na^+ or K^+ (Table 3). However, relatively high T_m values were obtained for some of the G/A-mutants that we can't explain so far. Thus, the thermal denaturation and CD analyses were in perfect agreement: if a PG4 sequence folded into a G-quadruplex in the presence of either Na^+ or K^+, it was detectable with both methods.

Chapitre 2, Table 3 Thermal denaturation analysis

5' UTR		No salt	LiCl	NaCl	KCl
EBAG9	wt	n.a.	n.a.	51.4 ± 1.1	>90
	mut	58.0 ± 1.3	72.7 ± 2.1	72.2 ± 0.9	71.2 ± 1.1
FZD2	wt	65.1 ± 1.8	77.0 ± 1.9	79.3 ± 1.7	>90
	mut	58.8 ± 2.3	70.5 ± 0.1	68.0 ± 1.5	64.0 ± 0.1
BARHL1	wt	40.1 ± 1.0	44.2 ± 1.9	70.6 ± 3.1	>90
	mut	48.2 ± 0.6	64.7 ± 1.7	61.8 ± 2.7	61.2 ± 1.6
NCAM2	wt	67.6 ± 4.0	78.4 ± 1.4	>90	>90
	mut	68.9 ± 1.7	70.4 ± 1.4	73.8 ± 2.8	72.1 ± 0.2
THRA	wt	67.8 ± 1.4	76.0 ± 0.8	75.4 ± 1.1	>90
	mut	82.3 ± 1.1	82.4 ± 1.2	82.6 ± 2.2	81.5 ± 1.0
AASDHPPT	wt	65.5 ± 1.8	77.7 ± 0.6	75.1 ± 4.7	>90
	mut	52.4 ± 0.7	59.2 ± 1.3	61.5 ± 0.2	59.1 ± 0.5
	snp	59.9 ± 2.2	75.5 ± 3.7	75.2 ± 1.7	73.7 ± 0.9
TNFSF12 C/A	wt	46.4 ± 3.1	60.5 ± 0.2	56.9 ± 0.9	79.8 ± 0.6
	mut[a]	49.0 ± 1.2	59.4 ± 1.9	62.5 ± 4.2	57.8 ± 1.9
DOC2B C/A	wt	59.8 ± 2.0	65.6 ± 0.5	68.4 ± 0.1	74.1 ± 0.5
	mut[b]	56.9 ± 0.4	64.3 ± 1.9	64.5 ± 0.1	63.3 ± 2.6
MAP3K11 C/A	wt	59.1 ± 0.5	68.1 ± 1.7	67.8 ± 0.4	73.9 ± 0.3
	mut[c]	54.0 ± 0.7	63.6 ± 0.1	64.1 ± 1.4	59.4 ± 2.0

n.a. The magnitude of the curves did not permit the determination of accurate T_m values
[a]Corresponds to the *TNFSF12* CG/AA-mutant version
[b]Corresponds to the *DOC2B* CG/AA-mutant version
[c]Corresponds to the *MAP3K11* CG/AA-mutant version

CD and thermal denaturation analyses are typical methods used to study G-quadrupex structures. However, because of their requirement for a relatively large amount of RNA (i.e. in the low micromolar range), they do not permit discrimination between the formation of a unimolecular, a bimolecular or a tetramolecular G-quadruplex structure. In the context of a G-quadruplex present in the 5'-UTR of an mRNA, the unimolecular topology is most likely; however, it is not impossible that several mRNAs may interact together through the formation of either a bimolecular or a tetramolecular structure. In order to address this question, an in-line probing was performed on all of the PG4 wild type and G/A-mutant versions. Trace amounts of 5'-^{32}P-radiolabelled transcripts were incubated for 40 hours in a slightly basic buffer (pH 8.3) that included a relatively high magnesium concentration (20 mM $MgCl_2$), and either in the absence or the presence of monovalent cations (Li^+, K^+ or Na^+). During the incubation, the presence of the magnesium led to the cleavage of the phosphodiester backbone of the single-stranded nucleotides often found at the periphery of the RNA structure [45]. If a PG4 sequence adopts a unimolecular G-quadruplex structure, the nucleotides in the loops should bulge out of the RNA's structure and therefore be susceptible to in-line attack by the magnesium ions. A typical example of an autoradiogram for an in-line attack experiment is illustrated for the *EBAG9* PG4-derived sequence in panel D of Figure 1. Clearly, an important difference in the intensity of the banding patterns was observed at several positions of the wild type PG4 in the presence of 100 mM KCl when compared to all other conditions. Specifically, there was a drastic increase in the intensity of the bands representing the nucleotides located between the guanosine tracks (e.g. C_{20}, A_{25} and A_{31}), and those corresponding to the loops of the PG4. In addition, the inability of the G/A-mutants to fold into a G-quadruplex structure was confirmed, regardless of the PG4 candidate. In order to provide a reliable evaluation, a quantitative analysis was performed. Briefly, at least two gels for each candidate were exposed to a phosphor screen and revealed by phosphor imaging using a Storm apparatus coupled with the SAFA software for the quantitative analysis [46]. The intensity of each band in the K^+ lane was divided by that of the corresponding band in the Li^+ lane. A nucleotide was considered

significantly more accessible when this ratio was higher than an arbitrarily fixed threshold of 2. A summary of all of the accessible nucleotides is shown in panel A of Figure 1. The nucleotides for which the accessibility was significantly modified by the addition of KCl are underlined. These nucleotides were always found to be located between the G-tracts, as well as in the vicinity of the PG4 (which should become single-stranded upon the formation of the G-quadruplex). These results validate the hypothesis that the G-quadruplex structures identified *in vitro* are able to fold according to a unimolecular topology. The conditions used in this experiment (i.e. trace amount of RNA, <1 nM) cannot trigger the formation of intermolecular G-quadruplexes. It is important to note that, to validate this technique, two different controls have been performed in conjunction to the in-line probing experiment for the *EBAG9* PG4-derived sequence. Firstly, the impact of a high concentration of magnesium (10 mM) on the G-quadruplex's formation was tested by CD experiments and it doesn't interfere with the ability to form a G-quadruplex structure (data not shown). Secondly, DMS probing experiment was perform in parallel of the in-line probing and many guanines in the tracks identified by bioinformatic were protected only in presence of 100 mM KCl (data not shown). Finally, the in-line probing data were in perfect agreement with that obtained from both the CD and the thermal denaturation analyses. The same set of six candidates identified in the previous experiments gave a positive G-quadruplex signature in the in-line probing experiments performed in the presence of KCl, while the other three did not. Moreover, only the *BARHL1* and *NCAM2* PG4-derived sequences appeared to fold into a G-quadruplex structure in the presence of NaCl.

In summary, the three different methods used provided consistent data for the set of PG4 candidates tested (see Table 4 for a summary). The PG4 sequences from the *EBAG9*, *FZD2*, *BARHL1*, *NCAM2*, *THRA* and *AASDHPPT* 5'-UTRs fold into G-quadruplex structures *in vitro* at a physiological concentration of KCl, while their G/A-mutant versions do not. The *TNFSF12*-, *MAP3K11*- and

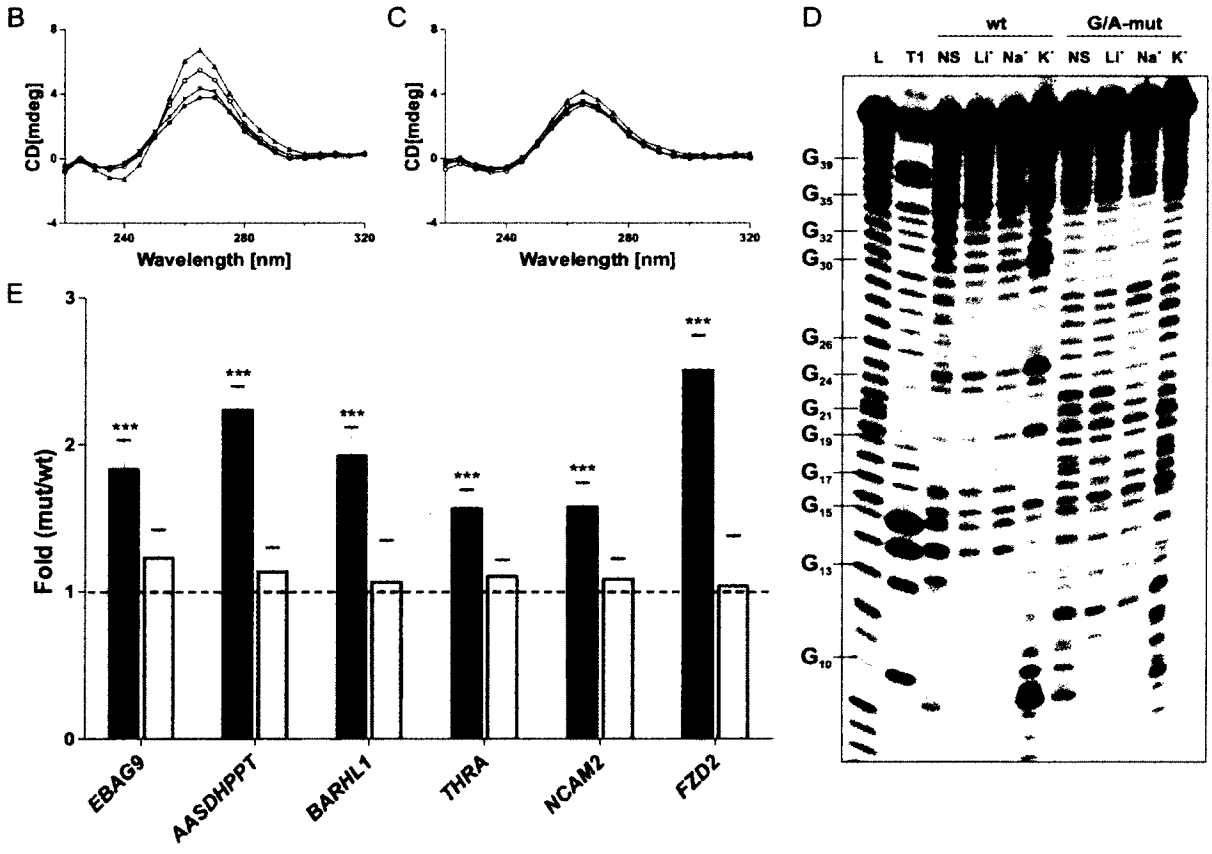
DOC2B-derived sequences do not fold into G-quadruplex structures under these conditions.

Ability of the identified G-quadruplexes to repress translation *in cellulo*

Subsequently, the characterization of the G-quadruplexes identified *in vitro* was performed by verifying their potential effects on translation *in cellulo*. Both the full-length wild type and the G/A-mutant 5'-UTRs of the candidates folding into G-quadruplexes *in vitro* were cloned upstream of a luciferase reporter gene (*Rluc*) (see Materials and Methods). HEK293 cells were then cotransfected with either the wild type or the G/A-mutated *Rluc* construction and a *Fluc* reporter gene, thereby permitting the normalization of the transfection efficiency. Cells were harvested 24 hours post-transfection and lysed. The resulting lysates were used in luciferase activity assays in order to estimate the quantity of luciferase protein synthesized. The Rluc activity was normalized with the Fluc activity for each sample. A ratio of luciferase activities was calculated by dividing the value determined for the G/A-mutant 5'-UTRs by that of the corresponding wild type 5'-UTR. This analysis yielded an estimation of the relative differences in luciferase protein resulting from the abolition of the G-quadruplex structure in each case. For example, the *EBAG9* G/A-mutated 5'-UTR construct (in which only 6 guanosines out of a total of 235 nucleotides were substituted for adenosines) produced a 1.8-fold greater level of luciferase activity than did its corresponding wild type counterpart (Figure 1E). The six G-quadruplex structures studied yielded estimated differences ranging from an increase of 1.56- to 2.50- fold in terms of the quantity of luciferase protein. In other words, the formation of the G-quadruplex structure significantly decreased the level of luciferase expression in all cases.

A

EBAG9	5' GGAGCCUCCGCCGGGCggGCggGGAGggGGAGGGGCAGGUUUUGA 3'
AASDHPPT	5' GgGgGCGgGCUGgGAGgGCUGUCGGUGGGCCAGUCUGC 3'
BARHL1	5' GGGGCAGCGGCGGCUGggGUUGgGgGUGGGUGggGAGCUUUUGGGG 3'
THRA	5' GGGUGCUGUGCCCUAGGGCCUGGGUGGCAgggggUgggUGGCCUGUGGG 3'
NCAM2	5' GGAGGAGCGCGCGgGCUGCGgGCGGCUGggGCACCGCGgGAGCGGCGGCGGCGG 3'
FZD2	5' GGGGAAGAAGCGCAGUCUCCGGGUUGGggGCGgGgGCGggGGggGCGCCAAGGAGCCGGG 3'
TNFSF12	5' GGCUCCCCCUCCCCCGAUCCCUCGgGUCCCGgGAUGgGGggGCGGUGAGGCAGG 3'
MAP3K11	5' GGCUCCCCAGAGAGGCGUGgGUCUGGgGCUGAGgGCCAGgGCCCGGAUGCCCAGG 3'
DOC2B	5' GGGCCUGCUGAGCCGCCCCCGgGCCGggGUCGCGCCGgGCCGgGCCGCGCCCGGGG 3'



Chapitre 2, Figure 1. Identification of G-quadruplex structures and translational repressors.

(**A**) Sequences of the wild type PG4s of the nine selected candidates. Sequences of the original potential G-quadruplexes are shaded in gray. The lowercase guanosines (g) correspond to those changed to adenosines in the G/A-mutants. Nucleotides that were cleaved significantly more in the presence of KCl, as compared to in the presence of LiCl, in the in-line probing experiments as determined by the quantification with the SAFA software, are underlined. (**B**) and (**C**) Circular dichroism spectra for the *EBAG9* PG4 candidate using 4 µM of either

the wild type (**B**) or the G/A-mutant (**C**) sequences and either in the absence of salt (close circles) or in the presence of 100 mM LiCl (closed triangles), NaCl (open circles) or KCl (open triangles). (**D**) Autoradiogram of a 10% denaturing (8 M urea) polyacrylamide gel of the in-line probing of the 5'-labelled *EBAG9* wild type and G/A-mutant PG4 versions performed either in the absence of salt (NS), or in the presence of either 100 mM LiCl, NaCl or KCl. The lanes designated L and T1 are an alkaline hydrolysis and a ribonuclease T1 (RNase T1) mapping of the wild type version, respectively. Representative guanosine residues are indicated on the left of the gel. (**E**) Gene expression levels of the different constructs used at the protein level using either the luciferase assay (black bars) or the mRNA level as determined by RT-qPCR (gray bars). The X axis identifies the candidates and the Y axis the fold difference that corresponds to the value obtained for the G/A-mutant version divided by that obtained for the wild type version for each candidate. The RT-qPCR values were obtained using the $\Delta\Delta C_T$ method with the *Fluc* gene as internal control and the wild type version as the calibrator. Error bars were calculated using a minimum of 3 independent experiments. *** indicates a *P*-value < 0.001.

Chapitre 2, Table 4. Summary of the *in vitro* and *in cellulo* analysis of the candidates in terms of their ability to adopt a G-quadruplex structure.

5' UTR	*In vitro*			*In cellulo* (Fold)
	CD	In line probing	T_m	
EBAG9	Yes	Yes	Yes	1.83
FZD2	Yes	Yes	Yes	2.50
BARHL1	Yes	Yes	Yes	1.92
NCAM2	Yes	Yes	Yes	1.57
THRA	Yes	Yes	Yes	1.56
AASDHPPT	Yes	Yes	Yes	2.24[a]/1.48[b]
TNFSF12 C/A	Yes	Yes	Yes	0.38[c]/1.29[d]
MAP3K11 C/A	Yes	Yes	-	-
DOC2B C/A	Yes	Yes	-	-

[a] Fold difference in protein expression for the *AASDHPPT* G/A-mutant *versus* the wt version
[b] Fold difference in protein expression for the *AASDHPPT* C7 SNP *versus* the wt version
[c] Fold difference in protein expression for the *TNFSF12* C/A-mutant *versus* the wt version
[d] Fold difference in protein expression for the *TNFSF12* CG/AA-mutant *versus* the wt version

RNA was also extracted from the above cells and RT-qPCR experiments performed in order to verify if the differences in gene expression occurred at the transcriptional or the post-transcriptional level. This analysis provided an evaluation of the quantity of mRNA produced by each construct. The same normalization methodology as was used with both the Fluc gene and the G/A-mutant version was utilized. The *EBAG9* 5'-UTRs produced the same Rluc mRNA level as did both the wild type and the G/A-mutant, namely ~1 (Figure 1E). Similar data were obtained for all of the other candidates. Because the mRNA levels did not vary between the wild type and the G/A-mutant versions, regardless of the candidate examined, this indicated that the formation of the G-quadruplex structure has a post-transcriptional effect.

The use of a different cell line yielded similar results at both the protein and the RNA levels (i.e. MCF-7 cells, data not shown). Similar experiments, but in which the reporter and normalizer genes were inverted (i.e. 5'-UTR inserted upstream of the Fluc gene), were also carried out and virtually identical data were obtained (data not shown). Together, these results show that all of the PG4 sequences able to fold into G-quadruplexes *in vitro* repressed the expression levels of two different reporter genes *in cellulo*, and did so in two different cell lines. Moreover, this repression occurs post-transcriptionally, most likely by repressing the translation level of the mRNA species in question.

Transforming negative candidates into positive candidates

Three of the nine candidates identified by the bioinformatic analysis were shown to be unable to fold into G-quadruplex structures in the presence of KCl (i.e. *TNFSF12*, *MAP3K11* and *DOC2B*). Since these sequences possessed the requirement of four consecutive guanosine tracts, we wondered why they did not adopt a G-quadruplex structure. Initially, the primary sequences of all of the PG4 sequences used for the *in vitro* experiments were compared. The first observation made was that these three sequences include significantly more cytosines than did

those that folded into G-quadruplexes (Table 5). In fact, the only interesting correlation observed was that a high G/C ratio appeared to be associated with the ability of the PG4 sequence to fold into a G-quadruplex structure (Table 5). The presence of a larger number of cytosines obviously lowers the G/C ratio. The relatively high level of cytosines most likely increases the ability of a given sequence to form stable stem structures resulting from GC Watson-Crick base pair formation. In order to verify this hypothesis, the stabilities, in terms of Gibbs free energy (ΔG) of the predicted secondary structures adopted by all of the PG4 sequences, were estimated using several bioinformatic programs (i.e. mfold, KineFold and MC-Fold) [47,48,49,50]. The predicted structures of the *TNFSF12*, *MAP3K11* and *DOC2B* sequences all had lower ΔG values, as compared to the others, regardless of the software used, indicating that they represent the most stable structures (Table 5). In these predicted structures, several of the guanosines required for the G-quadruplex formation where in fact involved in GC Watson-Crick base pairs. This has the effect of stabilizing the rod-like predicted secondary structure (data not shown). As is well known that the rod-like secondary structure is formed relatively rapidly, while G-quadruplex formation usually requires a fairly long period of time [44,51,52], we hypothesized that the presence of a relatively stable secondary structure may prevent the formation of a G-quadruplex of this is indeed the case, the reduction of the stability of the initial secondary structure should favor the formation of an alternative one that includes a G-quadruplex. Consequently, mutants in which several randomly chosen cytosines were substituted for adenosines were synthesized (i.e. C/A-mutants) (Figure 2A). The number of substitutions was calculated so as to yield a final G/C ratio equal to that of the lowest G/C ratio of the positive candidates, specifically the 2.2 of *NCAM2* (Table 5). Moreover, mutants in which important guanosines of the PG4 were mutated to adenosines, in addition to the C/A mutations, were also generated (CG/AA-mutants). In order to verify if these mutants were able to fold into G-quadruplex structures, they were subjected to the *in vitro* experiments described above.

Chapitre 2, Table 5. Primary sequence and secondary structure analysis of *in vitro* PG4s.

Gene	Length (nt)	Nb G	%G	Nb C	%C	%GC	%AT	Ratio G/C	mfold (kcal/mole)	KineFOLD (kcal/mole)	MC-Fold (kcal/mole)
EBAG9	45	26	57.8	9	18.4	77.8	22.2	**2.9**	-13.9	-14.2	-37.4
FZD2	60	36	60.0	13	21.7	80.0	20.0	**3.0**	-16.8	-18.6	-49.2
BARHL1	46	30	65.2	5	10.9	76.1	23.9	**6.0**	-8.5	-8.6	-30
NCAM2	54	33	61.1	15	27.8	88.9	11.1	**2.2**	-22.4	-21.1	-48.0
THRA	49	28	57.1	9	18.4	75.5	24.5	**3.1**	-19.7	-19.1	-38.4
AASDHPPT	38	22	57.9	8	21.1	79	21.0	**2.8**	-13.7	-14.2	-34.2
[illegible]											
TNFSF12 C/A	54	23	42.6	10	18.5	61.1	38.9	**2.3**	-10.2	-13.1	-42.5
[illegible]											
MAP3K11 C/A	55	26	47.3	10	18.2	65.5	34.5	**2.6**	-12.3	-14.5	-47.8
[illegible]											
DOC2B C/A	56	28	50.0	12	21.4	71.4	28.57	**2.3**	-9.7	-10.0	-43.5

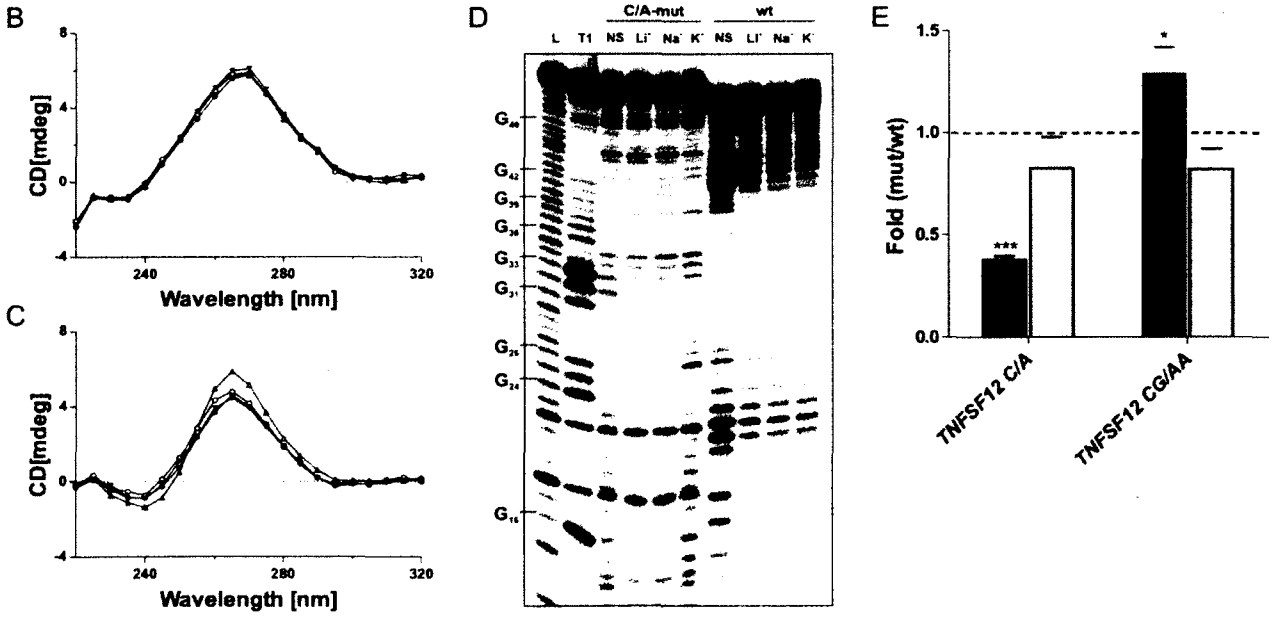
First, CD spectra experiments were performed on all of the C/A-mutants. The C/A-mutant of *TNFSF12* produced a shift to the G-quadruplex characteristic spectrum in the presence of KCl, while no change was observed for the corresponding wild type sequence (Figure 2B-C). Similar results were obtained for the C/A-mutants of both *MAP3K11* and *DOC2B* (Table 4). Second, thermal denaturation experiments corroborated the CD results. Specifically, all of the C/A-mutants showed a significant increase in the T_m value in the presence of KCl, while those of the CG/AA-mutants remained relatively constant. The increase in stability observed in the presence of KCl is a good indication that G-quadruplexes structures were adopted by the C/A-mutants. Finally, the in-line probing experiments also added to the physical evidence that the C/A-mutants fold into G-quadruplexes while their wild-type counterparts adopt rod-like structures involving GC Watson-Crick base pairs. For example, the wild type *TNFSF12* sequence's probing gel showed that the cytosine-rich sequence located from positions 5 to 13 most likely interacted with the guanosine-rich sequence located in positions 31 to 39 (Figure 2D). This helical region includes 7 GC, 1 AU and 1 GU base pairs out of a total of 18 nucleotides from both strands. Therefore, it appears reasonable to suggest that its formation impaired the G-quadruplex formation. In the case of the corresponding C/A-mutant, stronger bands corresponding to the loop appeared in

the presence of KCl only between the guanosine tracts, and in the middle of the 7 guanosine long tract (Figure 2D). As observed before, this pattern is characteristic of a G-quadruplex structure. Similar data were also obtained for the *MAP3K11* and *DOC2B* candidates. Clearly, the in-line probing experiments support the initial hypothesis that the stable secondary structures formed by these sequences prevent the formation of the G-quadruplexes. Together, these approaches make a strong case for explaining why, even though the *TNFSF12*, *MAP3K11* and *DOC2B* sequences possess all of the basic requirements for the adoption of a G-quadruplex structure in the presence of the KCl, they instead fold into a stable secondary structure containing a relatively long double-stranded helical domain. However, it should be noted that the introduction of mutations that destabilize this initial secondary structure favors the formation of the corresponding G-quadruplex structures.

A ***TNFSF12 C/A*** 5' GGaUCaCaCUCaCaCGAUaaaUaGGGUCCCGGGAUGGGGGGGGaGGUGAGGCAGG 3'

MAP3K11 C/A 5' GGaUCaCCAGAGAGGCGUGGGUCUGGGGaUGAGGGaCAGGGaCCGGAUGCCaAGG 3'

DOC2B C/A 5' GGGaaUGaUGAGaaGCCCaCGGGaCGGGGUCGaGaaGGGCCGGGaCGCGCaCGGGG 3'



Chapitre 2, Figure 2. Rescue of G-quadruplex structures *in vitro* and *in cellulo*.

(**A**) Sequences of the C/A-mutant PG4 versions for each of the 3 candidates that did not initially fold into G-quadruplex structures. The original PG4 sequences are highlighted in gray. Lowercase adenosines (a) were cytosines in the wild type changed to adenosines in the C/A-mutant versions. Lowercase guanosines (g) correspond to those changed to adenosines in the G/A-mutant versions. Underlined nucleotides correspond to those that were cleaved significantly more in the presence of KCl, as compared to in the presence of LiCl in the in-line probing experiment after quantification with the SAFA software. (**B**) and (**C**) Circular dichroism spectra for the *TNFSF12* candidate using 4 µM of either the wild type (**B**) or the C/A-mutant versions (**C**) and either in the absence of salt (close circles) or in the presence of either 100 mM LiCl (close triangles), NaCl (open circles) or KCl (open triangles). (**D**) Autoradiogram of a 10% denaturing (8 M urea) polyacrylamide gel of in-line probing of the 5'-end labeled *TNFSF12* C/A-mutant and wild type PG4 versions performed either in the absence of salt (NS) or in the presence of either 100 mM LiCl, NaCl or KCl. The lanes designated L and T1 are an alkaline hydrolysis and an RNase T1 mapping of the C/A-mutant version, respectively. Representative guanosine residues are indicated on the left of the gel. (**E**) Gene expression levels of the different constructs used at either the protein level, determined using luciferase assays (black bars), or the mRNA level, determined using RT-qPCR (gray bars). The X axis identifies the mutated version of the *TNFSF12* candidates, and the Y axis the fold difference corresponding to the value obtained for either the C/A-mutant or the CG/AA-mutant version divided by the value obtained for the wild type version of *TNFSF12*. The RT-qPCR values were obtained using the $\Delta\Delta C_T$ method with the *Fluc* gene as internal control and the wild type version as calibrator. Error bars were calculated using a minimum of 3 independent experiments. * indicates *P*-value < 0.05 and *** a *P*-value < 0.001.

Subsequently, whether or not this G-quadruplex rescue (i.e. the C/A-mutants) had the ability to repress translation *in cellulo* was investigated. The appropriate plasmid constructions (i.e. full-length wild type, C/A-mutant and CG/AA-mutant 5'UTR versions for *TNFSF12*) were cloned upstream of the *Rluc* reporter gene, transfected into HEK293 cells and the gene expression analyzed at both the protein and mRNA levels as described previously. Astonishing decreases in the amounts of protein synthesized for the C/A-mutants, as compared to those for the wild type versions, were observed (Figure 2E and Table 4). Specifically, in the case of the C/A-mutant of *TNFSF12*, a 2.6-fold less level of protein was detected. The CG/AA-mutant showed a small increase of 1.29-fold, at the protein level, giving a net repression estimated to be 3.3-fold by the *TNFSF12* G-

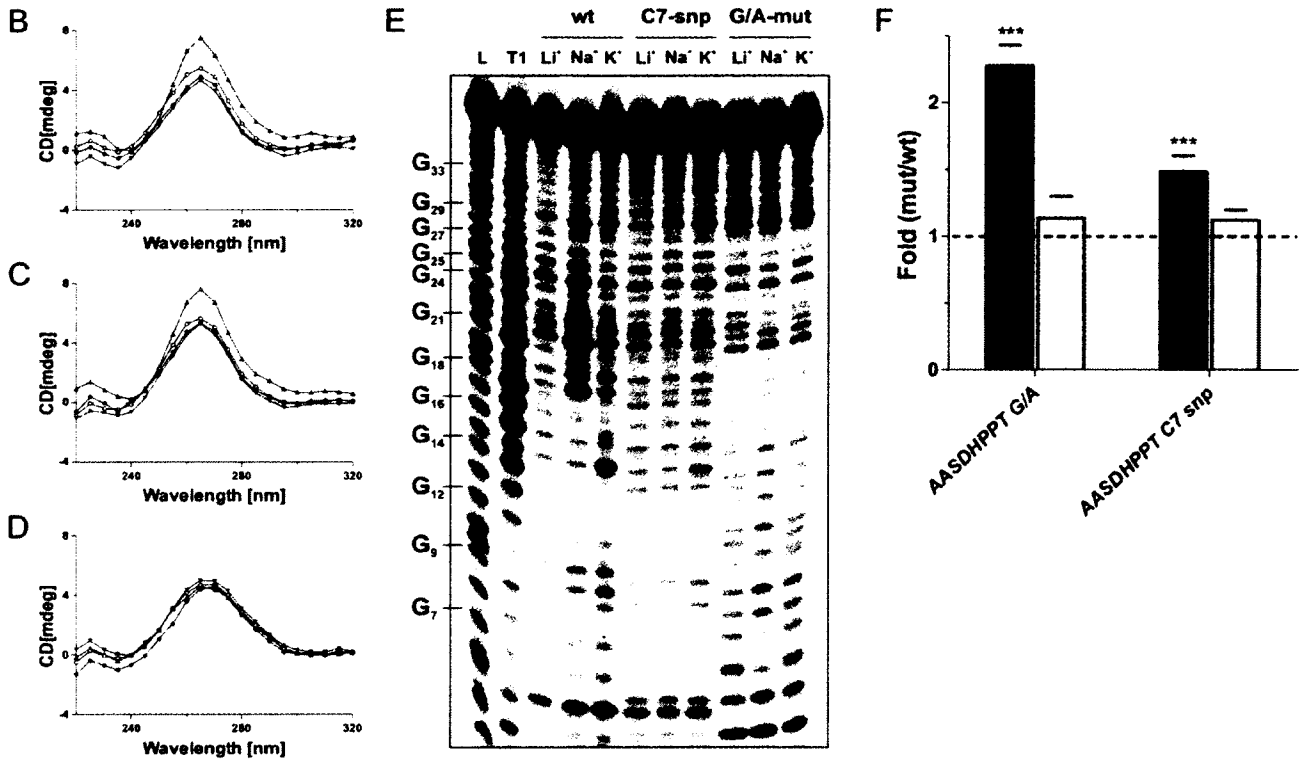
quadruplex. In all the cases, the mRNA level remained the same (Figure 2E). Thus, these results confirmed that by modifying the initial secondary structure, it was possible to modulate the formation of the G-quadruplex *in vitro* as well as *in cellulo*.

Single Nucleotide Polymorphisms in 5'-UTR G-quadruplexes.

According to the results presented here, it appears reasonable to suggest that G-quadruplexes located in the 5'-UTRs of mRNAs act as translational repressors of several genes in human cells. Therefore, it is logical to wonder if variability exists in these repressors, and, if yes, can this variability change the level of repression between individuals. A bioinformatic search was performed in order to identify Single Nucleotide Polymorphisms (SNP) within the human PG4 sequences from the UTRef collection of the UTRdb database (Dataset S3). A total of 327 SNPs were found in 271 different PG4 sequences with a distribution of 184 SNPs in 155 PG4 sequences located in the template strand, and 143 SNPs in 116 PG4 sequences located in the complementary strand (see Dataset S4). Thus, 5.0% of all PG4 sequences included at least one SNP. The PG4 with the highest number of SNPs was found in the 5'-UTR of the dihydrofolate reductase mRNA at position 35 (RefSeq: NM_000791). It contains a total of 8 different SNPs.

Interestingly, a SNP was identified in one of our initial candidates, namely *AASDHPPT*. It consisted of a substitution for the guanosine located in position 7 by a cytosine (Figure 3A). This guanosine was previously shown to be important for the formation of the G-quadruplex structure by the in-line probing analysis (Figure 1A). In order to investigate if this single substitution found in some individuals can affect not only the formation of the G-quadruplex structure, but also its ability to repress translation, the same set of *in vitro* and *in cellulo* experiments as described previously were performed. CD spectra analysis showed that both the wild type (G7) and SNP (C7) PG4 versions exhibited a G-quadruplex signature in the

A ***AASDHPPT C7*** 5' GGGGGCcGGCUGGGAGGGCUGUCGGUGGGCCAGUCUGC 3'



Chapitre 2, Figure 3. Effects of a SNP in a 5'-UTR G-quadruplex.

(**A**) Sequence of the C7 SNP PG4 version of the *AASDHPPT* candidate. The gray box indicates the sequence of the original PG4 identified by the algorithm. The lowercase cytosine (c) corresponds to the guanosine changed to a cytosine in the C7 SNP version. The underlined nucleotides correspond to the nucleotides that were cleaved significantly more in the presence of KCl, as compared to in the presence of LiCl, in the in-line probing experiment after quantification with the SAFA software. (**B**), (**C**) and (**D**) Circular dichroism spectra for the *AASDHPPT* candidate using 4 µM of either the wild type (**B**), the C7 SNP (**C**) or the G/A-mutant versions (**D**) and either in the absence of salt (close circles) or in the presence of 100 mM LiCl (close triangles), NaCl (open circles) or KCl (open triangles). (**E**) Autoradiogram of a 10% denaturing (8 M urea) polyacrylamide gel of the in-line probing of 5'-labelled wild type, C7 SNP and G/A-mutant PG4 versions of *AASDHPPT* performed either in the absence of salt (NS) or in the presence of 100 mM LiCl, NaCl or KCl. The lane designated L and T1 are an alkaline hydrolysis and an RNase T1 mapping of the wild type version, respectively. Representative guanosine residues are indicated on the left of the gel. (**F**) Gene expression levels of the different constructs used at the protein level, as determined using luciferase assays (black bars), or at the mRNA level, as determined using RT-qPCR (gray bars). The X axis identifies the mutated version of *AASDHPPT*, and the Y axis the

fold difference corresponding to the value obtained for either the G/A-mutant or C7 SNP version divided by the value obtained for the wild type version. The RT-qPCR values were obtained using the $\Delta\Delta C_T$ method with the *Fluc* gene as internal control and the wild type version as calibrator. Error bars were calculated using a minimum of 3 independent experiments. *** indicates *P*-value < 0.001.

presence of KCl, while the G/A-mutant did not (Figure 3, panels B-D). Conversely, the thermal denaturation experiment showed no increase in the T_m value in the presence of KCl for the SNP (C7), suggesting that the G-quadruplex structure was not adopted (Table 4). Similarly, the in-line probing gel displayed no specific structural rearrangement in the presence of KCl for the SNP (C7) version, result comparable to that observed for the G/A-mutant. It is noteworthy that the nucleotides located in the PG4 loop tended to become more accessible in the wild type (G7) sequence under these conditions (Figure 1A and Figure 3E). These banding patterns demonstrated that, in the presence of trace amounts of RNA, the wild type (G7) version can fold into a G-quadruplex structure *in vitro* in the presence of KCl while both the SNP (C7) and G/A-mutant versions did not. The discrepancy observed with the CD spectra may result from the large amount of RNA required by this method (i.e. micromolar quantities). This result suggests that CD analysis may provide misleading results.

Subsequently, the full-length wild type, SNP (C7) and G/A-mutant versions of the *AASDHPPT* 5'-UTR sequence were cloned upstream of the *Rluc* reporter gene. After the transfection of HEK293 cells, the levels of gene expression were monitored at both the protein and the mRNA levels by comparing either the G/A-mutant to the wild type (G7), or the SNP (C7) to the wild type (G7). Increases of 2.24- and 1.48-fold at the protein level were observed for the G/A-mutant and SNP (C7) sequences, respectively. At the mRNA level, no variation was observed in both cases (Figure 3F and Table 3), clearly showing that both the G/A-mutant and SNP (C7) versions were able to disrupt, or at least weaken sufficiently, the G-quadruplex structure leading to an increase in the translation of the downstream gene.

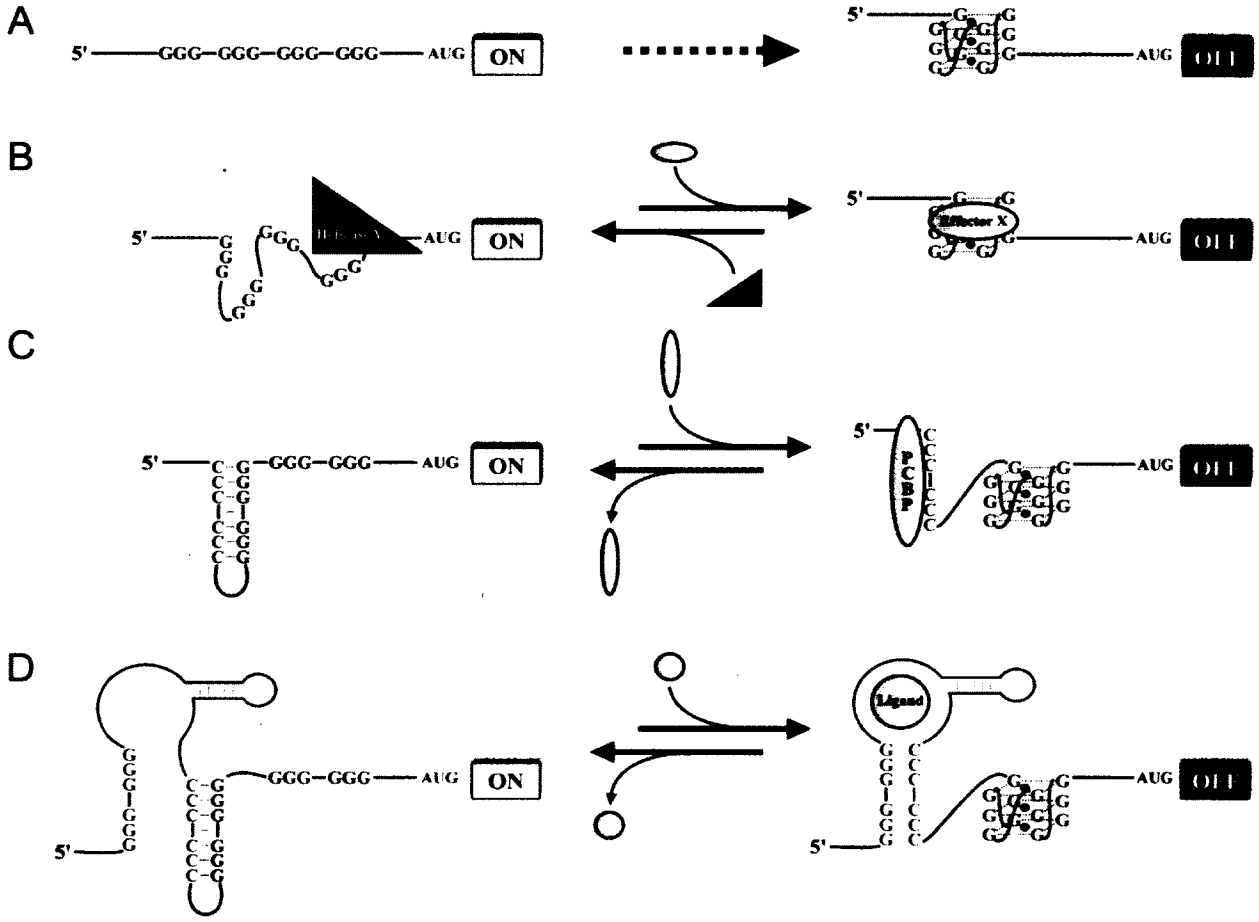
Discussion

The importance of the G-quadruplexes found in RNA molecules, in terms of the life cycle of the cell, remains to be appreciated. In fact, the field appears to be only in its infancy. Some researchers are investigating the physical rules that surround RNA G-quadruplex structures, while others try to find an associated biological role. The bioinformatic search reported here, as well as a previously reported one [12], demonstrate that sequences potentially capable of forming G-quadruplexes are located in thousands of 5'-UTRs. This observation led to the formulation of the hypothesis that G-quadruplexes are in fact translational repressors that are involved in various pathways within the cell. When analyzing the 5 024 different human mRNAs possessing at least 1 PG4 in their 5'-UTR retrieved in this work, we found not only for their presence noteworthy, but also the fact that they were found in a broad variety of genes in terms of gene ontology. For example, PG4 sequences were enriched in many of the mRNAs encoding the proteins involved in transcription regulation, mRNA transcription, protein modification, G-protein mediated signaling, cation transport and developmental processes, to name only a few examples (with *P*-values of 6.4×10^{-19}, 2.2×10^{-14}, 4.3×10^{-14}, 7.8×10^{-10}, 4.5×10^{-9} and 2.7×10^{-8}, respectively; see Dataset S5). However, it is important to consider that this type of bioinformatic search would undoubtedly overestimate the real prevalence and impact of G-quadruplex structures in the 5'-UTRs of the transciptome because it is based solely on sequence criteria. This point is well illustrated here by the fact that only six out of the nine PG4 candidates tested did in fact fold into a G-quadruplex structure (according to the *in vitro* experiments performed). However, if the resulting percentage of true G-quadruplex is indicative of all possibilities (67%), it suggests that there still are several thousand G-quadruplexes located exclusively in 5'-UTRs.

In order to obtain a reliable evaluation of the global importance of the presence of G-quadruplexes in the 5'-UTRs of mRNAs, we selected nine PG4 sequences retrieved in the mRNAs encoding proteins belonging to various cellular pathways. Of these, classical methods such as CD analysis and thermal denaturation (using a version smaller than the active 5'-UTR) provided consistent data indicating that six of the sequences were in fact folding into G-quadruplex structures (Figure 1 and Table 3). Specifically, the CD spectroscopy led to the observation of a G-quadruplex characteristic spectrum transition, while the thermal denaturation permitted the observation of higher T_m values in the presence of Na^+ or K^+, as compared to that expected for structures based on Watson-Crick base pairs. In order to provide additional physical support, in-line probing experiments were also performed. To our knowledge, this represents the first time that this technique was extensively used to analyze G-quadruplex structures, although it is routinely used to characterize other biologically relevant RNA structures such as riboswitches [45]. In-line probing is simple to perform, and does not require important RNA concentrations as compared to the other usual techniques. In addition, it is an efficient, reproducible and reliable method for studying RNA structure. The use of only trace amounts of RNA (<1 nM) most likely ensures analysis of the unimolecular structures, and strongly discourages intermolecular ones. This study provides data in agreement with the physico-chemical approaches, as well as permitting the determination of specific physical features such as the positions and the nucleotides of the loops of the G-quadruplex structures. All of the G-quadruplexes identified by in-line probing, possessed three loops intercalated by four guanosine tracks, confirming the formation of a unimolecular G-quadruplex. Moreover, the in-line probing gels permitted the determination of the nature of the inhibitory secondary structure formed by the three candidates that initially could not fold into a G-quadruplex. Finally, it was striking to observe that the six G-quadruplexes identified *in vitro* repressed translation *in cellulo* in the context of their full length 5'-UTR.

Taken together, the data from the *in vitro* and *in cellulo* experiments showed that the G-quadruplex structures are, indeed, very important in 5'-UTR sequences, specifically because of their ability to repress translation. In light of these results, they appear to be a key component in translational regulation (Figure 4A). Probably the easiest way to illustrate this is shown in Figure 4A where the simple presence of a guanosine-rich sequence in a 5'-UTR, in conjunction with the appropriate thermodynamic parameters, is sufficient to form a G-quadruplex structure. An interesting subsequent question to ask would be how this structure can be modulated in both space and time in the cell. G-quadruplex structures can certainly be the targets of specific proteins that decrease the minimal energy required for their formation; however, specific helicases have also been reported to unwind these stable RNA structures (Figure 4B)[26,53,54]. Alternatively, the three candidates that did not fold into G-quadruplex structures bring another level of complexity to the situation. Clearly, it is not simply because a 5'-UTR contains a PG4 sequence that it forms a G-quadruplex in the presence of K^+. The nature of the nucleotides located in the vicinity of the PG4 sequence is important in determining whether or not a G-quadruplex structure is adopted. In some cases, the G-quadruplex structure is not favoured over stable secondary structure based solely on Watson-Crick base pairs. The presence of cytosine tracks appears to be detrimental to G-quadruplex formation, as they interact and form stable stem secondary structures with the guanosine tracks. In this case, the driving forces involved in the G-quadruplex folding pathway will be too weak to promote its formation. Nevertheless, this characteristic could be implicated in the regulation of the formation of new G-quadruplexes. For instance, it increases the range of proteins potentially involved in these mechanisms to include poly(rC)-binding proteins, or stem-loop RNA helicases, that could disrupt the inhibitory secondary structure thereby allowing the G-quadruplex formation to proceed (Figure 4C). The GC stem secondary structures could also be disturbed by the RNA itself. As is observed for riboswitches, an RNA aptamer present either upstream or downstream of the G-quadruplex could change the local structure of the RNA upon the binding of a metabolite, thereby leading to the removal of the inhibitory GC

stem (Figure 4D). The G-quadruplex would act as the expression platform of such a riboswitch. With the RNA G-quadruplexes field growing rapidly, the discovery of more RNA G-quadruplex regulators will become essential to accurately defining their different roles.



Chapitre 2, Figure 4. Proposed models for the regulation by 5'-UTR G-quadruplexes.

Representations of different means of regulation by 5'-UTR G-quadruplexes. The general symbols are as follows: Green G, the guanosines involved in the G-quadruplex structure; AUGs, the initiation codon of the open reading frames; white (ON) and black (OFF) rectangles, reflect the translation status; and, small black spheres, represent the monovalent cation needed for the G-quadruplex formation (most likely K^+). (**A**) Simplest model for the G-quadruplex formation based solely on favorable thermodynamic parameters. (**B**) Modulation of the G-quadruplex formation by proteins interacting directly with the G-quadruplex structure. The black triangles represent a helicase with the specific activity of unwinding G-quadruplex

structures. Gray horizontal ovals represent a protein that binds G-quadruplex structures and either promotes its formation, and/or stabilizes the final structure. (**C**) The red C corresponds to the cytosine tracks that can interact, by the formation of Watson-Crick base pairs, with some of the guanosine tracks involved in the G-quadruplex structure. The gray vertical ovals represent a poly(rC)-binding protein (PCBP) that is able to bind the cytosine tracks and disrupt the initial inhibitory secondary structure, thereby allowing the formation of the G-quadruplex. (**D**) The red parts correspond to an RNA aptamer able to bind a specific ligand (gray spheres). The binding of the ligand promotes the final folding step of the aptamer and the formation of a new stem involving both the cytosine and guanosine tracks of the aptamer. This rearrangement removes the initial inhibitory secondary structure and allows the formation of the G-quadruplex.

Single nucleotide polymorphism (SNP) occurs when a single nucleotide in the genome differs between the members of a species. SNP are also involved in several diseases (e.g. cancer and Alzheimer's), and can be related to how a person will react to a specific treatment [55]. After having analyzed the impact of the flanking sequences of the G-quadruplex on their formation, the effect of a change in the sequence of the G-quadruplex itself was investigated. From all the bioinformatic results presented here, the database of SNPs inside 5'-UTR PG4s (Dataset S4) clearly represents the major novelty in the field concerning *in silico* information available to all and should be of general interest for researchers working in many fields. At least one SNP was found in 116 different 5'-UTR PG4s located on the complementary strand. Several of the corresponding genes are known to be implicated in various diseases (e.g. the *RAD51* (NM_002875) and *CAV2* (NM_001233) genes in cancer). The presence of the SNP within the *AASDHPPT* 5'-UTR, which was used as a model PG4 sequence, abolishes the G-quadruplex structure formation *in vitro* and increases the translation of a reporter gene *in cellulo*. These results suggest that two individuals could have a different expression for a given gene due to the difference in their PG4 SNP. Thus, SNPs located in 5'-UTR G-quadruplexes might be involved either in the predisposition, or in the appearance of, various diseases and cancers by altering the gene expression background of a specific individual. However, the bioinformatic approach used most likely identifies only the SNPs that lead to the abolition of a G-quadruplex because it searched for the presence of an SNP inside an already

discovered PG4. Moreover, it has been shown that the sequences adopting a non B-DNA structure (e.g. G-quadruplexes) possessed higher levels of polymorphism [56]. Consequently, there is a higher error frequency in these sequences during the DNA replication. This fact adds to the importance of exploring the presence of SNPs in G-quadruplexes. In summary, a deeper analysis of SNPs in G-quadruplex structures remains essential, but this study provides an original proof-of-principle of their relevancy.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the technical assistance of Patrice Coulombe.

FUNDING

This work was supported by a grant from the Canadian Institutes of Health Research (CIHR, grant number MOP-44022) to JPP. The RNA group is supported by grants from both the Université de Sherbrooke and the CIHR (grant number PRG-80169). JDB was the recipient of pre-doctoral fellowships from both the CIHR and the Fonds de Recherche en Santé du Québec (FRSQ). JPP holds the Canada Research Chair in Genomics and Catalytic RNA and is member of the Centre de Recherche Clinique Étienne-Lebel.

REFERENCES

1. ENCODE Project Consortium, Birney,E., Stamatoyannopoulos,JA., Dutta,A., Guigo,R. et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799-816.
2. Halbeisen,RE., Galgano,A., Scherrer,T. and Gerber,AP. (2008) Post-transcriptional gene regulation: from genome-wide studies to principles. *Cell. Mol. Life Sci.*, **65**, 798-813.
3. Ghildiyal,M. and Zamore,PD. (2009) Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.*, **10**, 94-108.
4. Bartel,DP. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **13**, 215-33.
5. Martick,M., Horan,LH., Noller,HF. and Scott,WG. (2008) A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. *Nature*, **454**, 899-902.
6. Roth,A. and Breaker,RR. (2009) The structural and functional diversity of metabolite-binding riboswitches. *Annu. Rev. Biochem.*, **78**, 305-34.
7. Neidle,S. and Balasubramanian,S. (2006) Quadruplex nucleic acids. *RSC Publishing*, Cambridge.
8. Burge,S., Parkinson,GN., Hazel,P., Todd,AK. and Neidle,S. (2006) Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.*, **34**, 5402-15.
9. Huppert,JL. (2008) Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. *Chem. Soc. Rev.*, **37**, 1375-84.
10. Verma,A., Halder,K., Halder,R., Yadav,VK., Rawal,P., Thakur,RK., Mohd,F., Sharma,A. and Chowdhury,S. (2008) Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J. Med. Chem.*, **51**, 5641-9.
11. Huppert,JL. and Balasubramanian,S. (2007) G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Res.*, **35**, 403-13.
12. Huppert,JL., Bugaut,A., Kumari,S. and Balasubramanian,S. (2008) G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, **36**, 6260-8.
13. Du,Z., Zhao,Y. and Li,N. (2009) Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic Acids Res.*, **37**, 6784-98.
14. Eddy,J. and Maizels,N. (2008) Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. *Nucleic Acids Res.*, **36**, 1321-33.

15. Parkinson,GN., Lee,MP. and Neidle,S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876-80.

16. Zaug,AJ., Podell,ER. and Cech,TR. (2005) Human POT1 disrupts telomeric G-quadruplexes allowing telomerase extension in vitro. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 10864-9.

17. Lipps,HJ. and Rhodes,D. (2009) G-quadruplex structures: in vivo evidence and function. *Trends Cell. Biol.*, **19**, 414-22.

18. Eddy,J. and Maizels,N. (2006) Gene function correlates with potential for G4 DNA formation in the human genome. *Nucleic Acids Res.*, **34**, 3887-96.

19. Siddiqui-Jain,A., Grand,CL., Bearss,DJ. and Hurley,LH. (2002) Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. *Proc. Natl. Acad. Sci. U.S.A.*, **99**, 11593-8.

20. Phan,AT., Kuryavyi,V., Burge,S., Neidle,S. and Patel,DJ. (2007) Structure of an unprecedented G-quadruplex scaffold in the human c-kit promoter. *J. Am. Chem. Soc.*, **129**, 4386-92.

21. Palumbo,SL., Memmott,RM., Uribe,DJ., Krotova-Khan,Y., Hurley,LH. and Ebbinghaus,SW. (2008) A novel G-quadruplex-forming GGA repeat region in the c-myb promoter is a critical regulator of promoter activity. *Nucleic Acids Res.*, **36**, 1755-69.

22. Paramasivam,M., Membrino,A., Cogoi,S., Fukuda,H., Nakagama,H. and Xodo,LE. (2009) Protein hnRNP A1 and its derivative Up1 unfold quadruplex DNA in the human KRAS promoter: implications for transcription. *Nucleic Acids Res.*, **37**, 2841-53.

23. Patel,DJ., Phan,AT. and Kuryavyi,V. (2007) Human telomere, oncogenic promoter and 5'-UTR G-quadruplexes: diverse higher order DNA and RNA targets for cancer therapeutics. *Nucleic Acids Res.*, **35**, 7429-55.

24. Saccà,B., Lacroix,L. and Mergny,JL. (2005) The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. *Nucleic Acids Res.*, **33**, 1182-92.

25. Kostadinov,R., Malhotra,N., Viotti,M., Shine,R., D'Antonio,L. and Bagga,P. (2006) GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. *Nucleic Acids Res.*, **34**, D119-24.

26. Darnell,JC., Jensen,KB., Jin,P., Brown,V., Warren,ST. and Darnell,RB. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. *Cell*, **107**, 489-99.

27. Gomez,D., Lemarteleur,T., Lacroix,L., Mailliet,P., Mergny,JL. and Riou,JF. (2004) Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. *Nucleic Acids Res.* **32**, 371-9.

28. Bonnal,S., Schaeffer,C., Créancier,L., Clamens,S., Moine,H., Prats,AC. and Vagner,S. (2003) A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons. *J. Biol. Chem.*, **278**, 39330-6.

29. Kumari,S., Bugaut,A., Huppert,JL. and Balasubramanian,S. (2007) An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.*, **3**, 218-21.

30. Arora,A., Dutkiewicz,M., Scaria,V., Hariharan,M., Maiti,S. and Kurreck,J. (2008) Inhibition of translation in living eukaryotics cells by an RNA G-quadruplex motif. *RNA*, **14**, 1290-6.

31. Morris,MJ. and Basu,S. (2009) An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells. *Biochemistry*, **48**, 5313-9.

32. Halder,K., Wieland,M. and Hartig,JS. (2009) Predictable suppression of gene expression by 5'-UTR-based RNA quadruplexes. *Nucleic Acids Res.*, **37**, 6811-7.

33. Mignone,F., Grillo,G., Licciulli,F., Iacono,M., Liuni,S., Kersey,PJ., Duarte,J., Saccone,C. and Pesole,G. (2005) UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.*, **33**, D141-6.

34. Jacobs,GH., Chen,A., Stevens,SG., Stockwell,PA., Black,MA., Tate,WP. and Brown,CM. (2009) Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, **37**, D72-6.

35. Macke,T., Ecker,D., Gutell,R., Gautheret,D., Case,DA. and Sampath,R. (2001) RNAMotif – A new RNA secondary structure definition and discovery algorithm. *Nucleic Acids Res.*, **29**, 4724-35.

36. Huang da,W., Sherman,BT. and Lempicki,RA. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44-57.

37. Beaudoin,JD. and Perreault,JP. (2008) Potassium ions modulate a G-quadruplex-ribozyme's activity. *RNA*, **14**, 1018-25.

38. Mergny,JL. and Lacroix,L. (2009) UV Melting of G-Quadruplexes. *Curr Protoc Nucleic Acid Chem.*, Chapter 17: Unit 17.1.

39. Livak,KJ. and Schmittgen,TD. (2001) Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods*, **25**, 402-8.

40. Todd,AK., Johnston,M. and Neidle,S. (2005) Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.*, **33**, 2901-7.

41. Huppert,JL. and Balasubramanian,S. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908-2916.

42. Tang,CF. and Shafer,RH. (2006) Engineering the quadruplex fold: nucleoside conformation determines both folding topology and molecularity in guanine quadruplexes. *J. Am. Chem. Soc.*, **128**, 5966-73.

43. Paramasivan,S., Rujan,I. and Bolton,PH. (2007) Circular dichroism of quadruplex DNAs: applications to structure, cation effects and ligand binding. *Methods*, **43**, 324-31.

44. Lane,AN., Chaires,JB., Gray,RD. and Trent,JO. (2008) Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.*, **36**, 5482-515.

45. Regulski,EE. and Breaker,RR. (2008) In-line probing analysis of riboswitches. *Methods Mol. Biol.*, **419**, 53-67.

46. Laederach,A., Das,R., Vicens,Q., Pearlman,SM., Brenowitz,M., Herschlag,D. and Altman,RB. (2008) Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments. *Nat. Protoc.*, **3**, 1395-401.

47. Mathews,DH., Sabina,J., Zuker,M. and Turner,DH. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911-940.

48. Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406-15.

49. Xayaphoummine,A., Bucher,T. and Isambert,H. (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acid Res.*, **33**, 605-610.

50. Parisien,M. and Major,F. (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, **452**, 51-55.

51. Onoa,B. and Tinoco,I Jr. (2004) RNA folding and unfolding. *Curr. Opin. Struct. Biol.*, **14**, 374-9.

52. Greenleaf,WJ., Frieda,KL., Foster,DA., Woodside,MT. and Block,SM. (2008) Direct observation of hierarchical folding in single riboswitch aptamers. *Science*, **319**, 630-3.

53. Wu,Y., Shin-ya,K. and Brosh,RM Jr. (2008) FANCJ helicase defective in Fanconia anemia and breast cancer unwinds G-quadruplex DNA to defend genomic stability. *Mol. Cell. Biol.*, **28**, 4116-28.

54. Creacy,SD., Routh,ED., Iwamoto,F., Nagamine,Y., Akman,SA. and Vaughn,JP. (2008) G4 resolvase 1 binds both DNA and RNA tetramolecular quadruplex with high affinity and is the major source of tetramolecular quadruplex G4-DNA and G4-RNA resolving activity in HeLa cell lysates. *J. Biol. Chem.*, **283**, 34626-34.

55. Shastry,BS. (2007) SNPs in disease gene mapping, medicinal drug development and evolution. *J. Hum. Genet.*, **52**, 871-80.

56. Wells,RD. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.*, **32**, 271-8.

Chapitre 2, Table S 1. Oligonucleotides used in this study.

For each candidate, the table provides both the names and the sequences of the oligonucleotides required for the synthesis of the different PG4 versions as well as for the construction of the different 5'-UTR versions.

Candidate	Name	5'- Sequence -3'
EBAG9	5UTR-1	GCGGGTTTCCCGATGAAGGGGCGGCCATGGCAGCTGCGCAGAGGCAACGCAGGCTGCTACGGAGCGCGCGCCCGGCTTTGAATG GGG
	5UTR-2	GCGCGCACACAAGGCGCGCTCACGCGCGCTGCCTGTCTCTTCCTCGAGCTCGCGCTCCGCCCGCTCACTCCCAGCCCCGCTCAT AAG
	5UTR-3 wt	GGTGGGAATCAAAACCTGCCCCTCCCCCTCCCCGCCCGCCCGGCGGAGGCTCCGAGCTGCCGCGGGCCGCGCGCGCACACAAGG
	5UTR-3 mut	GGTGGGAATCAAAACCTGCCCCTCCTTCTCCTTGCTTGCCCGGCGGAGGCTCCGAGCTGCCGCGGGCCGCGCGCGCACACAAGG
	5UTR-Forward	CATGCATGGCTAGCGCGGGTTTCCCGATGAAGGG
	5UTR-Reverse	GTACGTACGCTAGCGGTGGGAATCAAAACC
	PG4 wt	TCAAAACCTGCCCCTCCCCCTCCCCGCCCGCCCGGCGGAGGCTCCTATAGTGAGTCGTATTA
	PG4 G/Amut	TCAAAACCTGCCCCTCCTTCTCCTTGCTTGCCCGGCGGAGGCTCCTATAGTGAGTCGTATTA
THRA	5UTR Exon1-1	CATGCATGGCTAGCGAGGTGGCCTGCAGGGAGCAGCCGCGAGCCGGCCGGGCGCGCCGAGCCCGAGCCCCAGCCGGAG
	5UTR Exon1-2	GGGGGCGAGTGTGGCGGCCCCGCGGCTCCTCCGGCAGAGGCGGCGGCCGCTCTGGCTCCTCCCTCCCCCGCCCCGCTCCGGCTC CTC
	5UTR Exon1-3	GGGCCGCCACACTCGCCCCCCGCCCCCCCCCGCGCTCACTCGCACTCACACCCGGGCGCAGGAGGCGGGCGGCCCGGGCCCCACC CCC
	5UTR Exon1-4	CCTCGGCGTGGCGCGCCGGACCGGGCTGGGCAGCGACGCGGGGGTCTCAGCGCCCCGTGCTGGGGGCGTCCATGGGGGGCCGGT GCC
	5UTR Exon2-1	CGAGGGATCTCTGGACAGGACAAGACTCCGAAGCTACTCCCCCAGCACACAGCCCGGGACCCACAAACCCAGCTTGCCCCCAGC CCC
	5UTR Exon2-2	CAGCACTTGGCCTTTCACACCATGCCCTGCGCCCCAAGGGGGGCGGGCGGTGGGAGGGGCCAGGGAGTGGCAGGTGGGAGGGCT GGC
	5UTR Exon2-3 wt	AAAGGCCAAGTGCTGAGGCGGGTATCATGGGTGCTGTGCCCTAGGGCCTGGGTGGCAGGGGGTGGGTGGCCTGTGGGTGTGCCC GGG
	5UTR Exon2-3 mut	AAAGGCCAAGTGCTGAGGCGGGTATCATGGGTGCTGTGCCCTAGGGCCTGGGTGGCATAAAATAATTGGCCTGTGGGTGTGCCC GGG
	5UTR Exon2-4	GTACGTACGCTAGCTCACTTCAATTCCATCCAGGATGCCCTCCAGCACGCCAAGAGACTGGGGTGGGCACACTGGCCCCCCCGC ACC
	5UTR Exon1-Reverse	CGGAGTCTTGTCCTGTCCAGAGATCCCTCGGCGTGGCGCGCCGGACCGGG
	5UTR Exon2-Forward	CCCGGTCCGGCGCGCCACGCCGAGGGATCTCTGGACAGGACAAGACTCCG
	PG4 wt	CCCACAGGCCACCCACCCCCTGCCACCCAGGCCCTAGGGCACAGCACCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCCACAGGCCAATTATTTTATGCCACCCAGGCCCTAGGGCACAGCACCCTATAGTGAGTCGTATTA
BARHL1	5UTR-1	GCTAGCCTTTTGGATCTAATGCGCAGAGGAGGTTGGCCCAGAGCTCCCGGGCTCCCCCAAGGCTGAACTCCGTCCAAGGTGCC
	5UTR-2	CGCCGCTGCCCCTGCCCCTAGCTGCGGGCTGGCATGGGGAAGGCGGGCAGGGAGCCTGCGGGCACCTTGGACGGA
	5UTR-3 wt	GCTAGCAGCCAAGCTGCGACCTGTCCTCCCCAAAAGCTCCCCACCCACCCCCAACCCCAGCCGCCGCTGCCCCTGCCCCTAG
	5UTR mut Reverse	CCCCAAAAGCTCTTCACCCACTCTCAACTTCAGCCGCCGCTGCCCC
	5UTR mut Forward	GGGGCAGCGGCGGCTGAAGTTGAGAGTGGGTGAAGAGCTTTTGGGGAGGAC
	PG4 wt	CCCCAAAAGCTCCCCACCCACCCCCAACCCCAGCCGCCGCTGCCCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCCCAAAAGCTCTTCACCCACTCTCAACTTCAGCCGCCGCTGCCCCTATAGTGAGTCGTATTA
FZD2	5UTR-1	GCTAGCCGAGTAAAGTTTGCAAAGAGGCGCGGGAGGCGGCAGCCGCAGCGAGGAGGCGGCGGGGAAGAAGCGCAGTC
	5UTR-2 wt	GCTAGCGCTGGCCGCCGCCCCCCACCCGGCTCCTTGGCGCCCCCCCCGCCCCCGCCCCCAACCCGGAGACTGCGCTTCTTCC
	5UTR mut Reverse	CCCGGCTCCTTGGCGCTTCCTTCGCTCTCGCTTCCAACCCGGAGACTGC
	5UTR mut Forward	GAAGCGCAGTCTCCGGGTTGGAAGCGAGAGCGAAGGAAGCGCCAAGGAGCCGGG

	PG4 wt	CCCGGCTCCTTGGCGCCCCCCCCGCCCCCGCCCCCAACCCGGAGACTGCGCTTCTTCCCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCCGGCTCCTTGGCGCTTCCTTCGCTCTCGCTTCCAACCCGGAGACTGCGCTTCTTCCCCTATAGTGAGTCGTATTA
NCAM2	5UTR-1	GCTAGCGCTGCCGCCGCGGCGGCCGCTGCTGCTGCTGCTTCTGCCGCCGCTGCCGCCGCCGCTGCCTGGATATAGTGCGGCAAC GGAGCTTGCAGTC
	5UTR-2 wt	GCCCCGGCCGCCTCTGCTAGAGCCGCCGCCGCCGCTCCCGCGGTGCCCCAGCCGCCCGCAGCCCGCGCGCTCCTCCTCGCAAAC CTGCAAGCTCCGC
	5UTR-3	GCTAGCGTTCAGGACGTGACACCGGTGGACAGTTTCAAAGTTATTAAGTCCAGCCCTGGAGAGAGAACCTTTCGCTGCCCCGGC CTCTGC
	5UTR mut Reverse	CCGCCGCCGCCGCTCTCGCGGTGCTTCAGCCGCTCGCAGCTCGCGCGCTCCTCCTATAG
	5UTR mut Forward	GGAGGAGCGCGCGAGCTGCGAGCGGCTGAAGCACCGCGAGAGCGGCGGCGGCGGCTC
	PG4 wt	CCGCCGCCGCCGCTCCCGCGGTGCCCCAGCCGCCCGCAGCCCGCGCGCTCCTCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCGCCGCCGCCGCTCTCGCGGTGCTTCAGCCGCTCGCAGCTCGCGCGCTCCTCCTATAGTGAGTCGTATTA
AASDHPPT	5UTR-1 wt	GATCGATCGCTAGCGGGGGCGGGCTGGGAGGGCTGTCGGTGGGCCAGTCTGCGTAGCGAC
	5UTR-1 mut	GATCGATCGCTAGCGAGAGCGAGCTGAGAGAGCTGTCGGTGGGCCAGTCTGCGTAGCGAC
	5UTR-1 snp	GATCGATCGCTAGCGGGGGCCGGCTGGGAGGGCTGTCGGTGGGCCAGTCTGCGTAGCGAC
	5UTR-2	GGCCTGATGGCGCGGACGCAAACGTGGCCCCACCCCCTTCTTCCCGGCGTCCGTGCGCAGGGGACGGGCCGTCGCTACGCAGAC CCC
	5UTR-3	GATCGATCGCTAGCACACTGAAAGCGGACCTCGCCGCTATCTCGGGCCTGATGGCGCGGA
	PG4 wt	GCAGACTGGCCCACCGACAGCCCTCCCAGCCCGCCCCCTATAGTGAGTCGTATTA
	PG4 G/Amut	GCAGACTGGCCCACCGACAGCTCTCTCAGCTCGCTCTCTATAGTGAGTCGTATTA
	PG4 C7snp	GCAGACTGGCCCACCGACAGCCCTCCCAGCCGGCCCCCTATAGTGAGTCGTATTA
TNFSF12	5UTR-1 wt	GTCAGTCAGCTAGCCTCTCCCCGGCCCGATCCGCCCGCCGGCTCCCCCTCCCCCGATCCCTCGGGTCCC
	5UTR-2 wt	GTCAGTCAGCTAGCGGGGGCGGGGGGCTGTGCCTGCCTCACCGCCCCCCCATCCCGGGACCCGAGGGATC
	5UTR-1 G/Amut	GTCAGTCAGCTAGCCTCTCCCCGGCCCGATCCGCCCGCCGGCTCCCCCTCCCCCGATCCCTCGAGTCCC
	5UTR-2 G/Amut	GTCAGTCAGCTAGCGGGGGCGGGGGGCTGTGCCTGCCTCACCGCTTCCTCATCTCGGGACTCGAGGGATC
	5UTR-1 C/Amut	GTCAGTCAGCTAGCCTCTCCCCGGCCCGATCCGCCCGCCGGATCACACTCACACGATAAATAGGGTCCC
	5UTR-2 C/Amut	GTCAGTCAGCTAGCGGGGGCGGGGGGCTGTGCCTGCCTCACCTCCCCCCCATCCCGGGACCCTATTTATC
	5UTR-1 CG/AAmut	GTCAGTCAGCTAGCCTCTCCCCGGCCCGATCCGCCCGCCGGATCACACTCACACGATAAATAGAGTCCC
	5UTR-2 CG/AAmut	GTCAGTCAGCTAGCGGGGGCGGGGGGCTGTGCCTGCCTCACCTCTTCCTCATCTCGGGACTCTATTTATC
	PG4 wt	CCTGCCTCACCGCCCCCCCCATCCCGGGACCCGAGGGATCGGGGGAGGGGGAGCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCTGCCTCACCGCTTCCTCATCTCGGGACTCGAGGGATCGGGGGAGGGGGAGCCTATAGTGAGTCGTATTA
	PG4 C/Amut	CCTGCCTCACCTCCCCCCCCATCCCGGGACCCTATTTATCGTGTGAGTGTGATCCTATAGTGAGTCGTATTA
	PG4 CG/AAmut	CCTGCCTCACCTCTTCCTCATCTCGGGACTCTATTTATCGTGTGAGTGTGATCCTATAGTGAGTCGTATTA
MAP3K11	PG4 wt	CCTGGGCATCCGGGCCCTGGCCCTCAGCCCCAGACCCACGCCTCTCTGGGGAGCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCTGGGCATCCGGGCTCTGGCTCTCAGCTCCAGACTCACGCCTCTCTGGGGAGCCTATAGTGAGTCGTATTA
	PG4 C/Amut	CCTTGGCATCCGGTCCCTGTCCCTCATCCCCAGACCCACGCCTCTCTGGTGATCCTATAGTGAGTCGTATTA
	PG4 CG/AAmut	CCTTGGCATCCGGTCTCTGTCTCTCATCTCCAGACTCACGCCTCTCTGGTGATCCTATAGTGAGTCGTATTA
DOC2B	PG4 wt	CCCCGGGCGCGGCCCGGCCCGGCGCGACCCCGGCCCGGGGGCGGCTCAGCAGGCCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CCCCGGGCGCGGCTCGGCTCGGCGCGACTTCGGCTCGGGGGCGGCTCAGCAGGCCCTATAGTGAGTCGTATTA
	PG4 C/Amut	CCCCGTGCGCGTCCCGGCCCTTCTCGACCCCGTCCCGTGGGCTTCTCATCATTCCCTATAGTGAGTCGTATTA
	PG4 CG/AAmut	5'CCCCGTGCGCGTCTCGGCTCTTCTCGACTTCGTCTCGTGGGCTTCTCATCATTCCCTATAGTGAGTCGTATTA

SUPPLEMENTARY INFORMATION

Tous les Datasets peuvent être retrouvés dans le fichier .zip

Dataset S1 PG4 database obtained from Transterm. This database contains lists of the PG4s found in 18 different species, including humans. It provides information about the numbers of PG4 on each strand, the PG4 strand distribution and on the number of 5'-UTRs in the database for each species.

Dataset S2 PG4 database obtained from the UTRfull database.

Dataset S3 PG4 database obtained from the UTRef database.

Dataset S4 SNPs located inside the PG4 sequence database from the UTRef database.

Dataset S5 Gene ontology analysis results of the genes containing a PG4 located on the complementary strand of their 5'-UTRs from the UTRfull database.

Dataset S6 Gene list used for gene ontology analysis with DAVID web-accessible program.

CHAPITRE 3: Les G-quadruplexes présents dans les 3'-UTR des ARNm du transcriptome humain.

ARTICLE: Exploring mRNA 3'-UTR G-quadruplexes: evidence of roles in both alternative polyadenylation and mRNA shortening

Jean-Denis Beaudoin and Jean-Pierre Perreault

Article soumis et en révision chez: *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*)

AVANT-PROPOS:

J'ai conceptualisé le projet et réalisé 100% des expériences rapportées dans cet article. J'ai également monté toutes les figures et participé de façon très active à l'écriture du manuscrit.

RÉSUMÉ

Les séquences d'ARN riches en guanines peuvent se replier en strucutres tétrahélicoïdales non-canoniques appelées G-quadruplexes. Elles ont été démontrées pour être distribuées à travers tout le transcriptome chez les mammifères et pour représenter des éléments régulateurs clés dans plusieurs mécanismes cellulaires. Ceci dit, leur rôle dans le 3'-UTR des ARNm demeure à être élucidé et apprécié. Une analyse bioinformatique des 3'-UTR des ARNm a révélé un enrichissement en G-quadruplexes. Dans le but de faire la lumière sur le(s) rôle(s) de ces structures, celles retrouvées dans les gènes *LRP5* et *FXR1* ont été caractérisées *in vitro* et *in cellulo*. Ces G-quadruplexes dans ces 3'-UTR ont été démontrés pour augmenter l'efficacité de polyadénylation à des sites alternatifs, menant à l'expression de transcrits plus courts, et pour être capables d'interférer avec le réseau de régulation par les microARNs pour un ARNm spécifique. Clairement, les G-quadruplexes présents dans les 3'-UTR des ARNm sont des éléments de régulation en *cis* avec un impact significatif sur l'expression génique.

ABSTRACT

Guanine-rich RNA sequences can fold into non-canonical, four stranded helical structures called G-quadruplexes that have been shown to be widely distributed within the mammalian transcriptome as well as being key regulatory elements in various biological mechanisms. That said, their role within the 3'-UTR of mRNA remains to be elucidated and appreciated. A bioinformatic analysis of the 3'-UTRs of mRNAs revealed an enrichment in G-quadruplexes. In order to shed light on the role(s) of these structures, those found in the *LRP5* and *FXR1* genes were characterized both *in vitro* and *in cellulo*. The 3'-UTR G-quadruplexes were found to increase the efficiencies of alternative polyadenylation sites, leading to the expression of shorter transcripts, and to possess the ability to interfere with the miRNA regulatory network of a specific mRNA. Clearly, G-quadruplexes located in the 3'-UTRs of mRNAs are *cis*-regulatory elements that have a significant impact on gene expression.

INTRODUCTION

With the recent discovery that over 90% of the human genome is actively transcribed, the view of the transcriptome has completely changed[1]. Cells rely significantly on post-transcriptional regulation mechanisms in order to express a certain set of genes at a precise time, localization and magnitude. Therefore, exhaustive characterization of post-transcriptional regulatory elements is required for a better understanding of gene expression.

Guanine-rich nucleic acids have the ability to fold into a non-canonical four-stranded helical structure called a "G-quadruplex" (G4). In this structure, four co-planar guanines interact with one another through Hoogsteen base pairs and are stabilized by the presence of monovalent metal cations, usually potassium, that are stacked one over the other and form the core of the structure[2-4]. Genome-wide bioinformatic studies looking at the distribution of potential intramolecular G4, consisting of four consecutive runs of three of more guanines intercalated with connecting loop sequences, have been reported[5-7]. Enrichment in G4 motifs has been associated with telomeres, gene promoters, ribosomal DNA, recombination hotspots and both the 5'- and 3'-untranslated regions (UTRs) of mRNAs, suggesting a potential regulatory role for these structures in many processes[8-13]. G4 structures found in DNA have been the subject of considerable study; however, considering that the RNA version of this structure is generally more stable than its DNA counterpart, RNA should be more prone to fold into a G4 structure. The most studied RNA G4 structures are those located in the 5'-UTR of mRNA, which have been shown to be translational repressors[8,14-16]. A recent study also revealed that a G4 structure located in the 3'-UTR of two dendritic mRNAs can dictate their localization in neurites[17]. Moreover, RNA G4 structures have been reported to modulate the alternative splicing of the *TP53* gene (encoding the p53 protein) and the *hTERT* gene (encoding the telomerase reverse transcriptase)[18,19]. In the case of the *TP53* gene, an RNA G4 structure present downstream of the gene was reported to be crucial in maintaining an accurate 3'-end processing and function

under conditions of stressing DNA damage[20]. To date, this is the only reported indication that a RNA G4 structure located downstream of a gene may impact the polyadenylation process.

Polyadenylation is a fundamental processing step of mRNA maturation, and is essential for its export, stability and translation. The pre-mRNA is cleaved 10-30 nucleotides (nt) downstream of the polyadenylation (PA) signal (AAUAAA and its polymorphic variants) and then an untemplated poly-A stretch is added[21-23]. Most 3'-UTR also contain alternative polyadenylation (APA) signals[24], the use of which create the deletion of large portions of the 3'-UTRs as well as *cis*- and *trans*-acting regulatory elements. This 3'-UTR shortening may affect mRNA stability, translational efficiency, nuclear export and cytoplasmic localization[25]. For example, it has been reported that both the increase in stability and the translational efficiency of shorter mRNA isoforms is derived in part from the loss of microRNA-mediated repression[26]. A higher incidence of APA and 3'-UTR shortening was observed in cancer cells, suggesting a pervasive role for APA in oncogene activation without genetic alteration[27]. Clearly, a better understanding of the factors modulating APA is imperative.

Here, a robust approach including *in silico*, *in vitro* and *in cellulo* experiments that permitted the exploration of G4s located in human mRNA 3'-UTRs is presented. Specifically, two 3'-UTR G4s were studied in their different natural contexts, revealing several roles for these structures. Particular attention was focused on the modulation of APA by the G4 structure and on its impact on 3'-UTR mRNA shortening and gene expression.

RESULTS

Potential G-quadruplex sequences within human 3'-UTRs

A database of potential G-quadruplex (PG4) sequences located in the 3'-UTRs of known human mRNAs was constructed using the procedure described

previously[8,9]. PG4 sequences were identified on both strands using an algorithm that searches for the sequence G_x-N_{1-7}-G_x-N_{1-7}-G_x-N_{1-7}-G_x, where $x \geq 3$ and N is any nucleotide (A, C, G or U). The PG4s located on the template strands correspond to tracks of cytosines in the sequences database, while those located on the complementary strands, which can be found in mRNAs, are tracks of guanosines. The analysis was performed on the 33,694 3'-UTRs obtained from the UTRef collection (Table 1 and Supplementary Data Set S1; data obtained for the UTRfull collection can be found in Supplementary Data Set S2). A total of 8,903 PG4 sequences were retrieved in 5,046 (15.0%) 3'-UTRs. Each 3'-UTR contains at least one PG4, but may possess more. An unequal distribution of the PG4s between the two strands was observed (55.2% on the template DNA strand *versus* 44.8% on the complementary mRNA strand). A similar bias was observed in studies looking at the distribution of 5'-UTR PG4 sequences, and suggests potential biological repercussions[8,9]. The number of PG4 per 3'-UTR (ratio PG4/3'UTR) differs between the strands, with values of 1.55 for the template and 1.42 for the complementary strands (Table 1), respectively, suggesting that the cell is better able to deal with consecutive G4 structures within a given 3'-UTR on the DNA template strand than in the mRNA. Finally, the PG4 density in 3'-UTRs was estimated to be 0.130/kbase and 0.105/kbase for the template and complementary strands, respectively, which corresponds to a 2-fold enrichment as compared to the entire human genome (0.057/kbase using the same algorithm)[9].

A second bioinformatic analysis was performed in order to estimate the biological impact of these 3'-UTR PG4s. Gene ontology analysis revealed a significant enrichment in the number of PG4s in several categories of genes, including those involved in certain biological processes (e.g. neuron differentiation) and pathways (e.g. MAPK signaling pathway), to name two examples (see Table 2 and Supplementary Data Set S3). Moreover, analysis of the 3'-UTR PG4s in the OMIM database revealed that 308 of these mRNAs can be related to 573 different diseases, including cancer (see Supplementary Data Set S4). Thus, the 3'-UTR PG4 sequences are widely distributed within the transcriptome, and are potentially involved in various cellular mechanisms and diseases.

Chapitre 3, Table 1. Incidence of potential G-quadruplexes in a human 3'-UTR database.

	Template strand	Complementary strand	Total
Nb. of 3'UTR	-	-	33,694
Nb of 3'UTR with PG4 (%)	3,163 (9.4%)	2,794 (8.3%)	5,046 (15.0%)
3'UTR with 1 PG4 (%)	2,079 (65.7%)	1,973 (70.6%)	2,986 (59.2%)
3'UTR with more than 1 PG4 (%)	1,084 (34.3%)	821 (29.4%)	2,060 (40.8%)
Nb. PG4	4,917	3,986	8,903
% of PG4	55.2%	44.8%	-
Ratio PG4/3'UTR	1.55	1.42	1.76
PG4 density	0.130/kbase	0.105/kbase	0.235/kbase

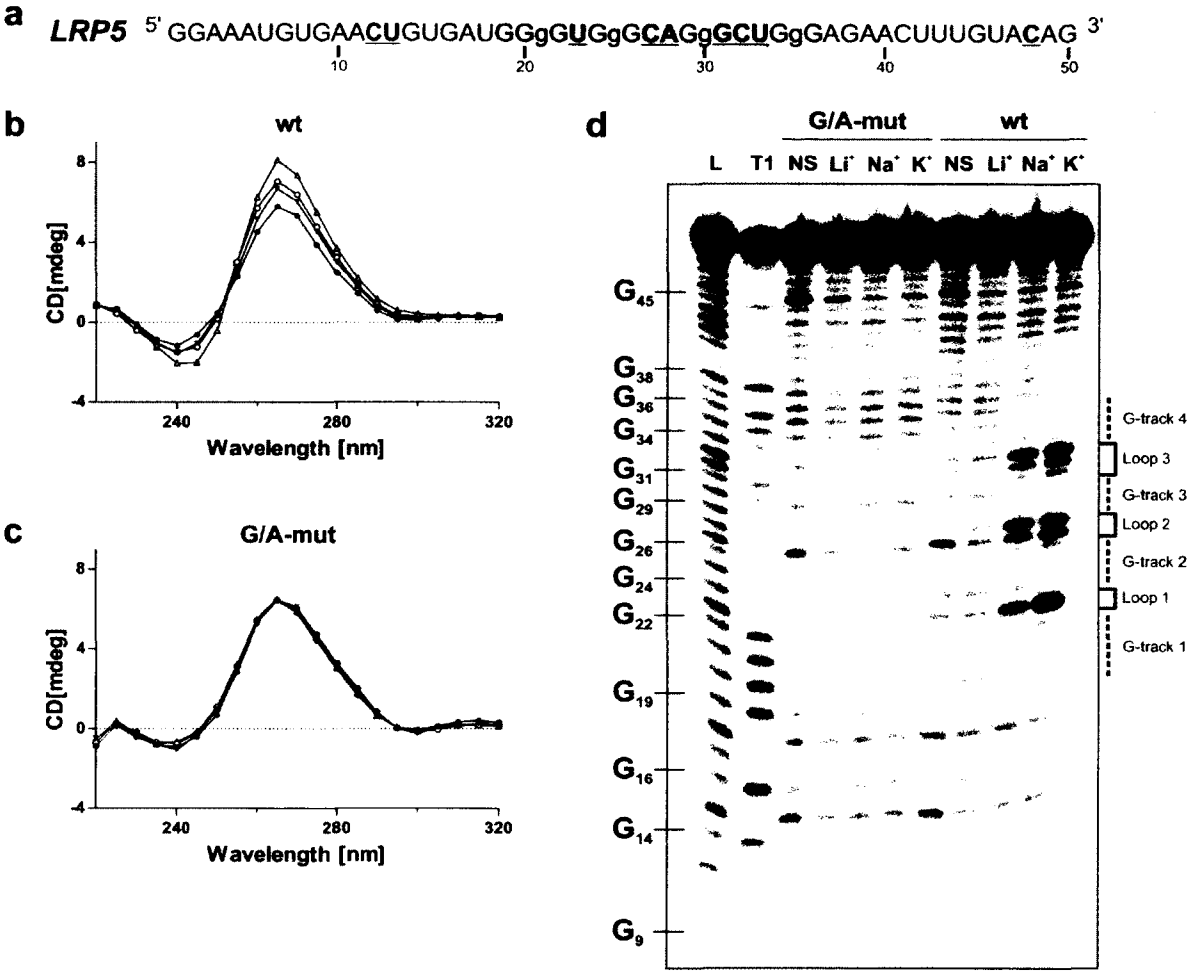
Chapitre 3, Table 2. Gene ontology analysis.

Categorie	Term	p-value
Biological process	Homophilic cell adhesion	2.90×10^{-06}
	Neuron differentiation	3.78×10^{-06}
	Cell adhesion	4.50×10^{-04}
	Regulation of neuron differentiation	9.73×10^{-04}
	Neuromuscular process	1.03×10^{-03}
Molecular function	Sequence-specific DNA binding	3.62×10^{-06}
	Calcium ion binding	5.18×10^{-04}
	Transcription factor activity	7.54×10^{-04}
	Chromatin binding	2.50×10^{-03}
	Transcription regulator activity	2.82×10^{-03}
	Protein kinase inhibitor activity	3.97×10^{-03}
Cellular component	Plasma membrane	5.47×10^{-05}
	Plasma membrane part	8.45×10^{-04}
	Synapse	5.38×10^{-03}
	Microtubule associated complex	8.54×10^{-03}
Pathway	MAPK signaling pathway	2.46×10^{-08}
	Notch signaling pathway	1.10×10^{-03}
	Arrhythmogenic right ventricular cardiomyopathy	3.45×10^{-03}
	Dilated cardiomyopathy	5.47×10^{-03}

In vitro testing of G-quadruplex formation in LRP5

Initially the PG4 located in the 3'-UTR of the low-density lipoprotein receptor-related protein 5 (*LRP5*) mRNA was studied as a model candidate. This PG4 sequence possesses small loops, a high number of guanosines and a low number of cystosines in flanking sequences; consequently, it possesses a strong predisposition to fold into a G4 structure (Fig. 1a). Moreover, the full length *LRP5* 3'-UTR is relatively short (203 nucleotides, nt), which significantly simplifies both the manipulations and the analysis of the data.

First, a sequence that exceeded the *LRP5* PG4 by ~15 nt at both ends was examined in order to evaluate its ability to fold into a G4 structure *in vitro* (Fig. 1a). A G/A-mutant version, created by the substitution of several guanosines for adenosines (i.e. to prevent the formation of a G4 structure), was also synthesized for use as a negative control. Firstly, G4 formation was monitored by circular dichroism (CD), a conventional method for which the four-stranded helical structures possess a typical spectrum. Due to the presence of the ribose residue, an RNA G4 structure is forced to adopt a parallel topology that is characterized by the appearances of both a negative peak at 240 nm and a positive one at 264 nm[28]. The CD spectra for both the wild-type (wt) (Fig. 1b) and G/A-mutated (Fig. 1c) versions were initially recorded either in the absence of salt, or in the presence of 100 mM LiCl (two conditions that do not support the formation of G4 structures), and then in the presence of 100 mM of either NaCl or KCl (two conditions that favor the formation of such structures). A significant transition through a characteristic parallel G4 structure was observed only for the wt version, especially in the presence of KCl. This supports the folding into a G4 structure within the *LRP5* 3'-UTR. Secondly, thermal denaturation analyses were performed. The formation of a G4 should lead to a significant increase in stability that is accompanied by a higher T_m value for the RNA species in question[29]. When the experiment was performed, significant increases in the T_m were only observed for the wt version in the presence of both NaCl and KCl (Table 3). The presence of



Chapitre 3, Figure 1. *LRP5* 3'-UTR PG4 folds into a G4 structure *in vitro*.

(**a**) Sequence and numbering of the wt *LRP5* PG4 used in the *in vitro* experiments. The lowercase guanosines (g) correspond to those mutated to adenosines in the G/A-mutant version. Nucleotides that were hydrolyzed significantly more in the presence of KCl during the in-line probing are both in bold and underlined. (**b,c**) Circular dichroism spectra for the *LRP5* PG4 sequence using 4 µM of either the wild-type (**b**) or the G/A-mutant (**c**) versions performed either in the absence of salt (close circles), or in the presence of 100 mM of either LiCl (close triangles), NaCl (open circles) or KCl (open triangles). (**d**) Autoradiogram of a 10% denaturing polyacrylamide gel of the in-line probing of the 5'-end-labeled *LRP5* wt and G/A-mutant PG4 versions performed either in the absence of salt (NS), or in the presence of 100 mM of either LiCl, NaCl or KCl. Lanes L and T1 correspond to alkaline hydrolysis and RNase T1 mapping of the wt version, respectively. The positions of the guanosines are indicated on the left of the gel, while the domains of the G4 structure are indicated on the right.

LiCl only induced a small increase in the T_m value due to stabilization of the RNA structure caused by a counter ion effect of the cations that attenuated the repulsion of the negatively charged phosphate backbone. Thirdly, in-line probing analyses, which require only trace amounts of RNA (<1 nM) favoring the formation of the G4 unimolecular topology that is most likely representative of that found within mRNAs[8], were performed. This method differs from both the CD and thermal denaturation methods which both require relatively large amounts of RNA (i.e. in the low µM range). During the incubation, the magnesium preferentially hydrolyzes the phosphodiester backbone of both flexible and single-stranded nucleotides such as those often found at the periphery of RNA structures[30]. Upon the formation of the G4 structure, the nucleotides located in the loops should bulge out and, therefore, be more susceptible to in-line attack by the magnesium ions. The *LRP5* PG4-derived sequences demonstrated this phenomenon. More specifically, the bands corresponding to the nucleotides located in the predicted loops, that is to say between the guanosine tracks (e.g. U_{23}, C_{27}, A_{28} and U_{33}), became drastically more intense only for the wt version in the presence of either NaCl or KCl (Fig. 1d). Quantifications' of the intensity of each band in presence of either LiCl or KCl indicated that the nucleotides that became more susceptible to hydrolysis in the presence of potassium were all proposed to be located in single-stranded regions within the G4 structure (Fig. 1a, bold and underlined nucleotides)[31]. All of the results obtained from the three distinct methods demonstrated that the *LRP5* PG4 sequence folds, *in vitro*, into a stable unimolecular G-quadruplex at physiological KCl concentrations (i.e. 100 mM).

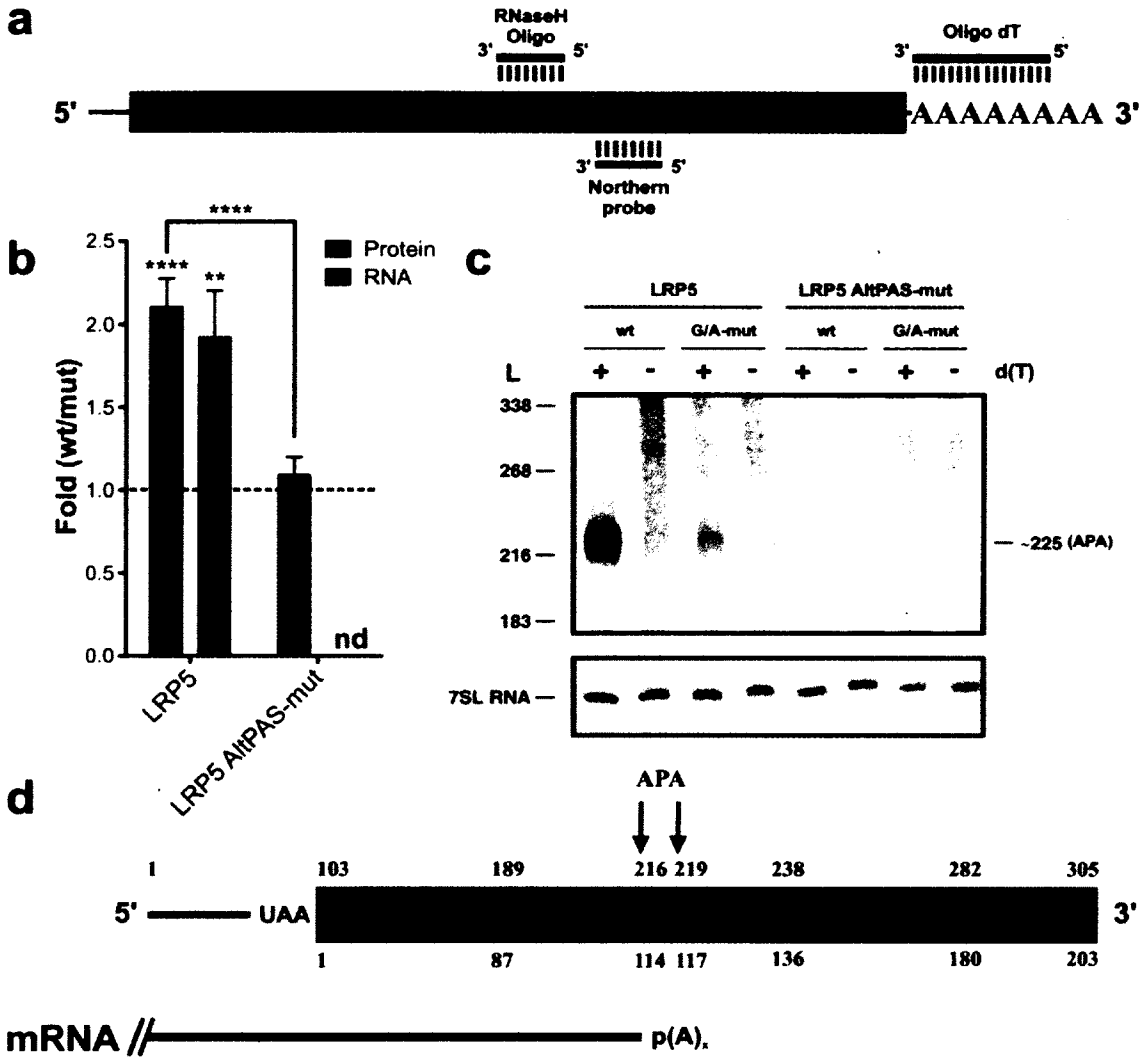
Chapitre 3, Table 3. Thermal denaturation analysis. Values shown are the means ± s.d. of two independent experiments.

3' UTR		No salt	Li+	Na+	K+
LRP5	wt	39.1 ± 2.1	51.3 ± 0.7	69.0 ± 0.6	>90
	mut	37.2 ± 1.9	49.0 ± 0.1	51.1 ± 1.8	47.3 ± 1.3
FXR1	wt	35.3 ± 0.6	57.1 ± 0.1	80.3 ± 0.1	>90
	mut	40.3 ± 0.1	52.1 ± 1.0	55.1 ± 0.9	48.8 ± 0.3

The LRP5 3'-UTR G-quadruplex influences gene expression *in cellulo*

The full-length *LRP5* 3'-UTR was cloned downstream of the firefly luciferase reporter gene (*Fluc*) in order to verify its ability to affect gene expression (Fig. 2a). During the cloning, the SV40 late polyadenylation signal in the pGL3 vector was removed from the construction. Only the natural polyadenylation signal (PAS) of the *LRP5* 3'-UTR, which is located 24 nt away from the polyadenylation site and corresponds to the polymorphic sequence UAUAAA was kept. HEK293T cells were then co-transfected by both Fluc-*LRP5* constructions (i.e. either with the wt or the G/A-mutant versions of the G4 structure) and a plasmid containing the renilla luciferase gene (*Rluc*) for normalization of the transfection efficiency. The cells were harvested 24 h post-transfection, lysed and luciferase activity assays performed in order to estimate gene expression. The ratio of the luciferase activities (value of the wt 3'-UTR divided by that of the G/A-mutant version) showed a 2-fold increase (Fig. 2b), indicating that the formation of the G4 structure significantly enhanced the luciferase expression level.

RNA samples were also extracted from the cells, and RNase H treatment coupled to Northern blot hybridization was performed in order to verify whether or not a correlation existed between the amounts of cellular proteins and mRNAs. Briefly, DNA oligonucleotides that specifically bind to a region 102 nt upstream of the *Fluc* gene's stop codon were annealed to the mRNA (see Fig. 2a) and the resulting RNA/DNA heteroduplex was then hydrolyzed by RNase H treatment. This removed the 5'-end of the Fluc-*LRP5* mRNAs, thereby permitting fractionation of the remaining 3'-ends by denaturing PAGE electrophoresis followed by Northern blot hybridization using a probe specific for the remaining part of the *Fluc* coding sequence regardless of the sequence of the 3'-UTR. The RNase H hydrolysis was performed in either the absence or the presence of oligo-d(T), which caused the heterogeneous polyadenylated products to collapse into discrete products. A single well-defined band was observed only in the presence of oligo-d(T) for both the wt and G/A-mutant versions, indicating that they correspond to polyadenylated RNAs (Fig. 2c). The wt version produced more mRNA as compared to the G/A-mutant,



Chapitre 3, Figure 2. The *LRP5* 3'-UTR G4 structure *in cellulo*.

(**a**) Schematic representation of the Fluc-*LRP5* construction. The *Fluc* coding sequence is shown in green, while the *LRP5* 3'-UTR is shown in blue. The binding regions of the oligonucleotides used for the RNase H hydrolysis, as well as the luciferase specific probe, are illustrated. (**b**) Gene expression levels of the different *LRP5* constructs either at the protein level (green), or the mRNA level (blue). The x-axis identifies the constructions used and the y-axis the fold difference (i.e. wt result divided by G/A-mutated result) (for LRP5 protein n = 3 while for mRNA n = 5, for *LRP5* AltPAS-mut protein n = 4, nd indicates not detectable). Error bars, mean ± s.d. ***P*<0.01 and *****P*<0.0001. (**c**) Northern blot hybridization of RNA samples subjected to a RNase H hydrolysis in the presence of a *Fluc* specific DNA oligonucleotide and either in the absence (-) or the presence (+) of oligo-dT. The numbers on the left refer to the sizes of a molecular ladder, while that on the right is the estimated size of the detected transcript. *7SL* RNA was probed as internal control. (**d**) Schematic view of the RNA product resulting from the RNase H

hydrolysis. The upper numbers correspond to the numbering from the 5' end of the digestion product, while lower ones refer to the start of the *LRP5* 3'-UTR. The arrows map the different polyadenylation sites as determined by 3'-RACE, and the mRNA produced is depicted in black.

although the abundance of *7SL* RNA (used as a loading control) remained invariable. The differences in the mRNA levels were in good agreement with what was observed at the protein level (Fig. 2b). A representation of the RNase H cleavage product for the Fluc-*LRP5* 3'-UTR is shown in Fig. 2d illustrating the 102 nucleotides from the RNase H cleavage site to the *Fluc* stop codon, the restriction site and the full-length *LRP5* 3'-UTR, which starts at position 103. The distance from the RNase H cleavage site to the *LRP5* polyadenylation site was estimated, by comparison to an RNA ladder, to be 220 nt, which was unexpected (see below). In order to confirm this evaluation, an 3'-RACE (rapid amplification of cDNA ends) experiment was performed, permitting resolution at the nucleotide level. Two very close, but distinct, polyadenylation sites were detected, generating fragments of 216 and 219 nt in size (i.e. corresponding to positions 114 and 117 of the *LRP5* 3'-UTR; Fig. 2d), thus validating the previous observation. These bands were not produced from the canonical polyadenylation site located at position 305 (according to NCBI), but instead from an alternative polyadenylation unit situated around positions 216 and 219 and under the control of an AAUAAA PAS located at position 189. This observation suggested that the G4 act as a downstream polyadenylation regulatory element that enhances the efficiency of the alternative polyadenylation unit, although it is excluded from the produced mature isoform. In order to test this hypothesis, new constructions possessing a mutated AAUAAA PAS (AltPAS-mut), which inactivates this APA unit, were synthesized for both the wt and the G/A-mutant G4 versions. No differences were observed in the luciferase activity levels, and no polyadenylation was detected in the *LRP5* 3'-UTR nor in its vicinity (Fig. 2b and c). The very low amount of luciferase protein produced, which was unaffected by either the presence or the absence of the G4 structure, potentially came from a PAS present in the pGL3 vector (located ~3,000 nt downstream the *LRP5* 3'-UTR) that was impossible to detect by the RNase

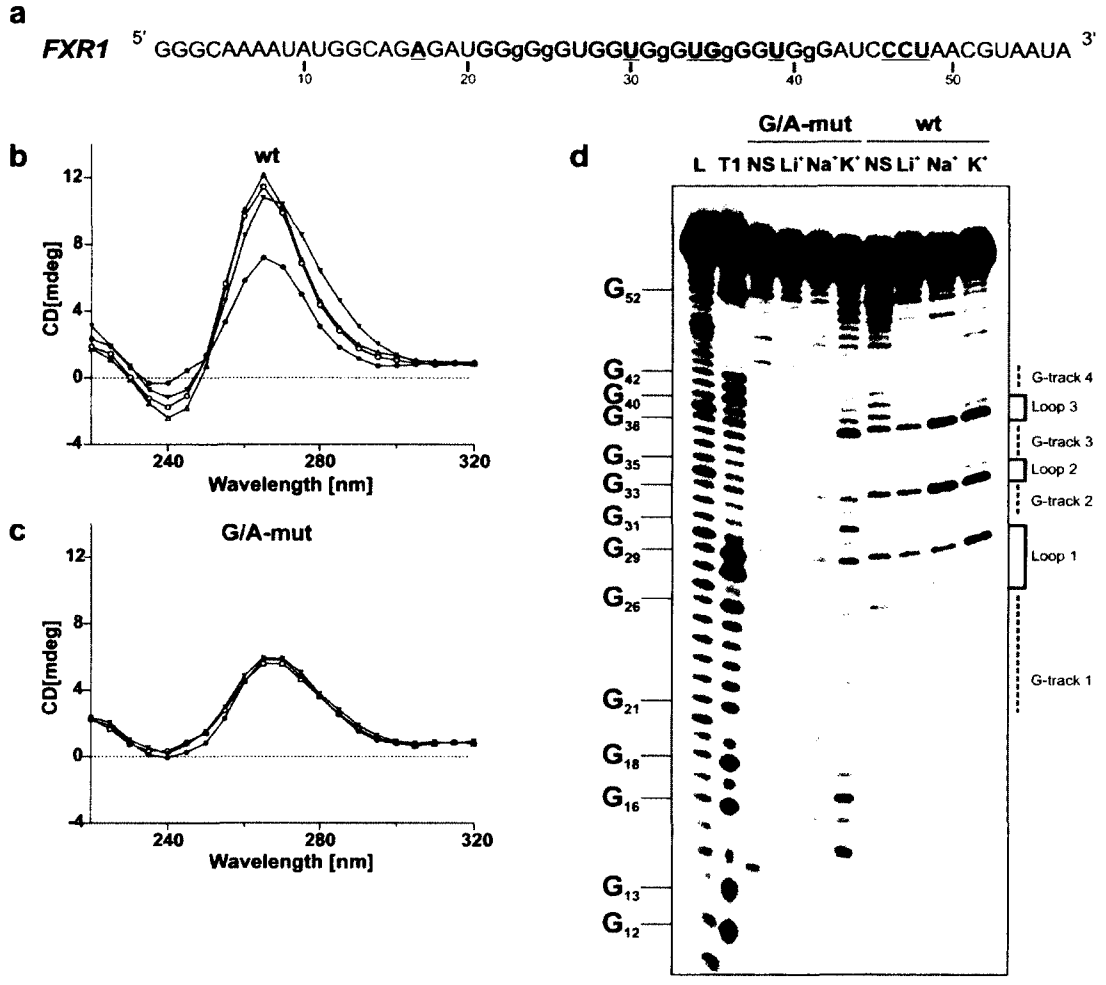
H/Northern blot experiment under the used conditions. Together, these results demonstrated that the G4 structure located within the LRP5 3'-UTR acts as a downstream regulatory element and that it positively modulates the use of an alternative polyadenylation unit.

3'-UTR G-quadruplexes appear to be frequently associated with alternative polyadenylation units

The human 3'-UTR mRNA database was revisited in order to identify PG4 sequences potentially involved in the regulation of an alternative polyadenylation unit. Each PG4 sequence was examined for the presence, within the first 100 nucleotides upstream (an arbitrarily chosen distance), of either a typical human PAS (i.e. AAUAAA) or the most common single polymorphism (i.e. AUUAAA). This analysis revealed the presence of 75 and 39 3'-UTR PG4s located near downstream AAUAAA and AUUAAA PAS, respectively, that formed putative alternative polyadenylation units that could potentially be linked to 22 different diseases (Supplementary Data Set 5). This suggests that the case of *LRP5* is not isolated, and that 3'-UTR G4 structures may be noteworthy *cis*-acting elements for the regulation of alternative polyadenylation.

A G-quadruplex structure promotes FXR1 3'-UTR shortening

In order to further evaluate the role of G-quadruplexes as positive regulatory elements for alternative polyadenylation units, a second candidate was studied. The fragile X related mental retardation autosomal homolog 1 (*FXR1*) gene produces an mRNA with a 3'-UTR 870 nt in length that possesses both a PG4 sequence and a putative internal alternative polyadenylation unit located around position 250 (Fig. 3a; note that the numbering from the positions of the *FXR1* 3'-UTR differs because the 102 upstream nucleotides of the *Fluc* coding sequence and the restriction site are also considered). Initially, the ability of the *FXR1* 3'-UTR PG4 sequence to fold into a G-quadruplex *in vitro* was assessed. The same three

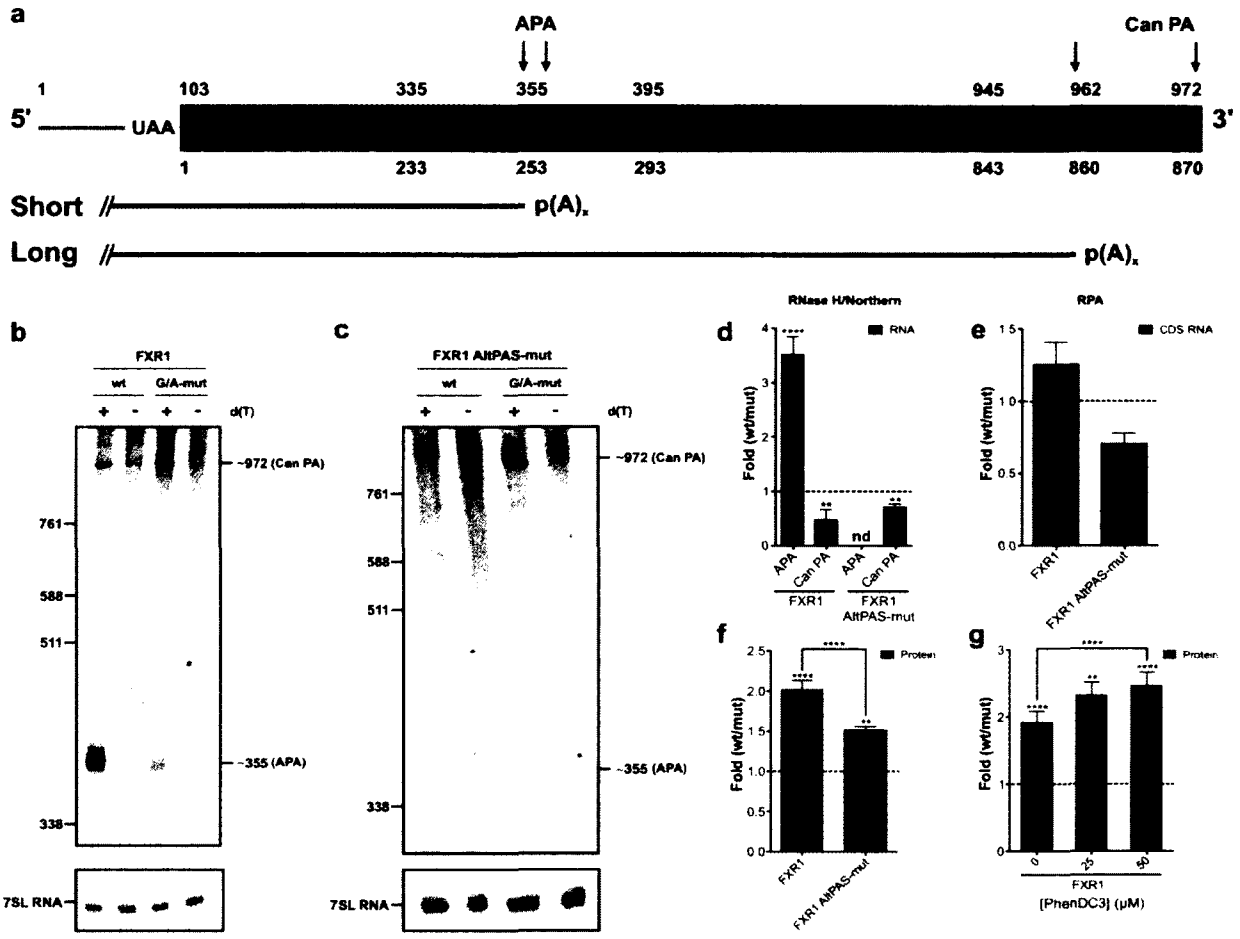


Chapitre 3, Figure S 1. *FXR1* 3'-UTR PG4 folds into G4 structure *in vitro*.

(**a**) Sequence and numbering of the wt *FXR1* PG4 used for the *in vitro* experiments. The lowercase guanosines (g) correspond to those mutated to adenosines in the G/A-mutant version. Nucleotides that were hydrolyzed significantly more in the presence of KCl during the in-line probing are both in bold and underlined. (**b,c**) Circular dichroism spectra for the *FXR1* PG4 sequence using 4 µM of either the wild-type (**b**) or the G/A-mutant (**c**) versions and either in the absence of salt (close circles) or in the presence of 100 of mM LiCl (close triangles), NaCl (open circles) or KCl (open triangles). (**d**) Autoradiogram of a 10% denaturing polyacrylamide gel of the in-line probing of the 5'-end-labeled *FXR1* wt and G/A-mutant PG4 versions performed either in the absence of salt (NS), or in the presence of 100 of mM LiCl, NaCl or KCl. Lanes L and T1 correspond, respectively, to alkaline hydrolysis and RNase T1 mapping of the wt version. The positions of the guanosines are indicated on the left of the gel, while the domains of the G4 structure are shown on the right.

methods described above were employed, and all agreed that it adopts a G4 structure in the presence of a physiological concentration of KCl (see Supplementary Results, Supplementary Fig. 1).

Subsequently, the full-length *FXR1* 3'-UTR was cloned downstream of the *Fluc* gene to verify its impact on gene expression. According to the primary sequence of the 3'-UTR, two mRNA species could be synthesized: one long isoform produced from the canonical polyadenylation site (AAUAAA PAS located 28 nt upstream of the predicted cleavage site); and, a shorter isoform produced from an alternative polyadenylation site (AUUAAA PAS located 60 nt upstream of the *FXR1* 3'-UTR G4 (Fig. 3a)). The *in cellulo* experiments were performed as described above. RNase H/Northern blot hybridization analysis confirmed the detection of both isoforms for both the wt and G/A-mutant versions only in the presence of oligo-d(T) (Fig. 3b). The shorter polyadenylated RNA species was estimated to be ~355 nt, and the longer one ~970 nt. These lengths correlated, respectively, with the positions of the alternative and canonical polyadenylation units. In agreement, 3'-RACE results indicated that the cleavage sites of both the alternative and the canonical polyadenylation units were located at positions 355-357 and 962-970, respectively (Fig. 3a). Quantification of the intensities of the bands for both isoforms revealed a 3-fold increase in the presence of the G4 structure for the shorter isoform, while in the longer isoform this decreased by 2-fold under the same condition (Fig. 3d). The *FXR1* G4 structure appears to significantly affect the short/long ratio of the produced mRNA isoforms. In order to investigate whether or not only the ratio between both isoforms was affected, or if the levels of total mRNA also varied, a second mRNA quantification was performed. An RNase protection assay (RPA) using probes that covered regions within the coding sequences of both the *Fluc* and *Rluc* genes (for normalization purposes) were performed. This approach permitted the quantification of the level of global mRNA synthesis without discriminating between the different isoforms. Almost no difference was observed, indicating that the *FXR1* G4 structure did not



Chapitre 3, Figure 3. The *FXR1* 3'-UTR G4 structure *in cellulo*.

(**a**) Schematic representation of the Fluc-*FXR1* transcripts resulting from the RNase H hydrolysis. The upper numbers correspond to the numbering from the 5' end of the hydrolyzed product, while lower ones refer to the start of the *FXR1* 3'-UTR (blue part). The arrows map the different polyadenylation sites as determined by the 3'-RACE experiments (alternative (APA) and canonical (Can PA) sites). The short and long mRNA isoforms produced are shown in black. (**b,c**) Northern blot hybridizations of the RNA samples previously subjected to RNase H hydrolysis in either the absence (-) or the presence (+) of oligo-dT. The numbers on the left refer to the sizes of a molecular ladder, while those on the right are the estimated sizes of the two isoforms. *7SL* RNA was probed as an internal control. (**d-f**) Gene expression levels of constructs either at the mRNA level as determined by Northern blot hybridization (for *FXR1* n = 5, while for *FXR1* AltPAS-mut n = 3; nd indicates not detectable) (**d**), by RNase protection assay (*FXR1* and *FXR1* AltPAS-mut n = 3) (**e**) (blue), or at the protein level as determined by luciferase assay (*FXR1* n = 7, FXR1 AltPAS-mut n = 3) (**f**) (green). The x-axis identifies the constructions used and the y-axis the fold difference (wt result divided by G/A-mutated result). (**g**) Luciferase assays in the presence of various concentrations of PhenDC3 (0-50 µM; n = 3). Error bars, mean ± s.d. ***P*<0.01 and *****P*<0.0001.

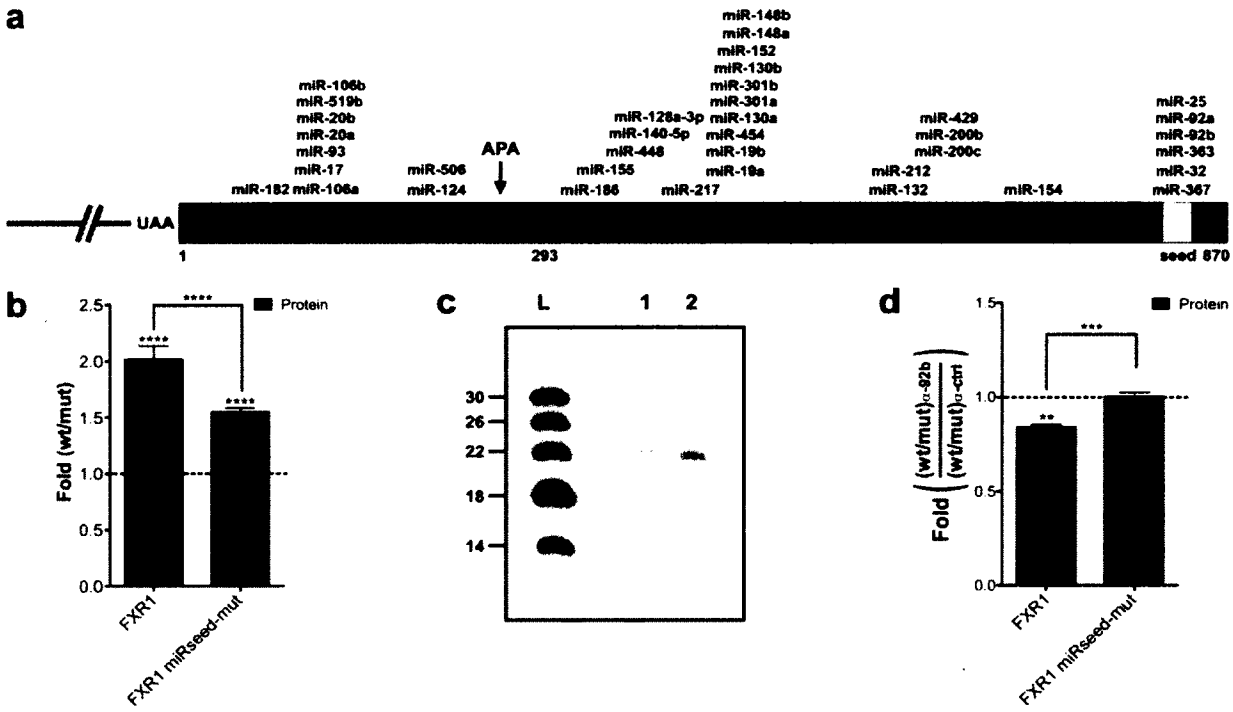
affect the global quantity of mRNA, but instead affected only the short/long isoform ratio (Fig. 3e). Interestingly, at the protein level, luciferase activity was increased by 2-fold in the presence of the *FXR1* G4 structure (Fig. 3f). These experiments demonstrated that the *FXR1* G4 structure influences gene expression at the protein level primarly by affecting the ratio between the short and the long mRNA isoforms without affecting the global mRNA level.

Afterwards, new constructions in which the AUUAAA PAS was mutated (AltPAS-mut) in both G4 contexts (i.e. for both the wt and G/A-mutant versions) were synthsized in order to verify whether or not the *FXR1* G4 structure positively modulates the efficiency of the alternative polyadenylation unit. The insertion of this mutation completely abolished the activity of the alternative polyadenylation unit and the synthesis of the shorter isoform (Fig. 3c). Interestingly, a significant decrease in the quantity of the long isoform was still observed in the presence of the G4 structure (Fig. 3c and 3d). Quantification of the mRNA produced at the canonical polyadenylation site, based on the RNaseH/Northern blot hybridization analysis, correlates with the total mRNA detected by the RNase protection assays for the alternative PAS mutants (Fig. 3d and 3e). At the same time, the luciferase activity assays showed a diminished increase (1.5-fold) with the inactive alternative polyadenylation unit constructions as compared to the active ones (2-fold) (Fig. 3f). This observation suggests that approximately half of the increase at the protein level in presence of the G4 structure is due to the stimulation of the alternative polyadenylation unit. Moreover, the effect of the G4 structure on the amount of mRNA synthesized at the downstream canonical polyadenylation site (30% decrease) seems likely to be independent of the utilization of the alternative site that is located upstream (Fig. 3d and 3e). Importantly, the experiment suggests that a smaller amount of mRNA harboring the *FXR1* G4 structure produced a larger amount of protein than did a larger amount of mRNA lacking the G4 structure. This represents an original characterization of this phenomenon.

In order to obtain support for the conclusion that the effects observed on gene expression were due to the presence of G4 structure, the impact of a G-quadruplex specific ligand on gene expression was tested. Specifically, PhenDC3

is a bisquinolinium derived compound with both a strong G4 stabilizing ability and selectivity[32,33]. The luciferase activity was observed to increase with increasing PhenDC3 concentrations, thus providing additional evidence that the *FXR1* G4 structure directly contributes to the differences observed in gene expression (Fig. 3g).

Finally, the impact of the *FXR1* 3'-UTR shortening was then investigated in terms of its microRNA regulatory network, which is also known as *trans*-factor elements regulating both mRNA stability and translation efficiency[34]. First, the mirSVR software was used to map the predicted microRNA binding sites present in the *FXR1* 3'-UTR (Fig. 4a)[35]. Only sites with a mirSVR score lower than -0.5 were considered. The loss of all of the predicted microRNA binding sites located downstream of the alternative polyadenylation unit during the 3'-UTR shortening process should likely lead to a modification of the microRNA-mediated regulation of the mRNA. The *FXR1* 3'-UTR has already been shown to be the target of various microRNAs, especially a seed region that is shared between 5 different microRNAs located at position 813 (Fig. 4a; yellow box)[36]. In order to test whether or not the increase in gene expression caused by the variation of the short/long isoform ratio driven by the *FXR1* G4 structure came from the loss of this negative regulatory element, constructions in which the conserved and shared seed region was mutated (*FXR1* miRseed-mut) were synthesized. The mutation of the seed region led to a reduction of the effect (50%) of the *FXR1* G4 as measured by luciferase activity (Fig. 4b). The same decrease was observed for *FXR1* AltPAS-mut, suggesting an important role for this region in gene expression due to the modulation of the APA site (Fig. 3f). Moreover, the microRNA miR-92b, which was proposed to bind to this seed, was detected by Northern blot hybridization in RNA samples from HEK293T cells (Fig. 4c). In order to enhance the role of miR-92b in this phenomenon, experiments using either a miR-92b inhibitor, or an irrelevant inhibitor control, were performed with the constructions. A decrease of more than 15% was observed for the natural *FXR1* context in the presence of the miR-92b specific inhibitor, while constructions harboring the miRseed-mutation remained



Chapitre 3, Figure 4. *FXR1* 3'-UTR shortening and the microRNAs regulatory network.

(**a**) Schematic representation of the *FXR1* 3'-UTR. The numbering refers to the position from the start site of the *FXR1* 3'-UTR. All predicted microRNA target sites with a mirSVR score < -0.5 according to the miRanda algorithm are shown[35]. The yellow region corresponds to the predicted shared microRNA seed region that was mutated in the FXR1 miRseed-mut constructions. (**b**) Gene expression levels of different *FXR1* constructs at the protein level as determined by luciferase assays. The x-axis identifies the constructions used and the y-axis the fold difference (wt result divided by G/A-mutated result) (for both *FXR1* and miRseed-mut n = 4). (**c**) Northern blot hybridization for the detection of miR-92b performed using either 5 µg (lane 1) of small RNAs (<200 nt) or 50 µg of total RNA (lane 2) extracted from untransfected HEK293T cells. The numbers on the left refer to the sizes of a molecular ladder of 5' end labeled *in vitro* transcripts (lane L). (**d**) Gene expression levels of different *FXR1* constructs at the protein level as determined by luciferase assays in the presence of either 100 nM miR-92b inhibitor or of irrelevant control inhibitors. The x-axis identifies the constructions used and the y-axis the fold difference (ratio wt on G/A-mutated version obtained in the presence of the miR-92b inhibitor divided by that obtained in presence of the control inhibitor) (both *FXR1* and miRseed-mut n = 3). ***P*<0.01, *** *P*<0.001 and *****P*<0.0001.

unaffected. These results support the hypothesis that most of the impact on gene expression caused by the *FXR1* 3'-UTR mRNA shortening promoted by the G4 structure comes from the modification of its microRNA regulatory network.

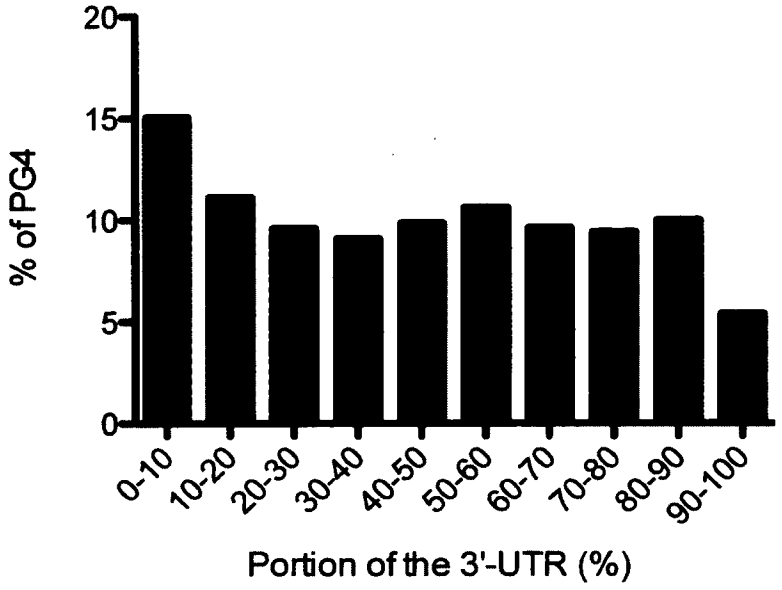
Discussion

In contrast to DNA G4 structures, the importance of both the presence and the impacts of RNA G4 structures in biology remains to be elucidated and appreciated. The bioinformatic analysis reported here is in agreement with a previous one showing that PG4 sequences are found in thousands of human 3'-UTRs[9], including in numerous mRNAs of proteins related to both human diseases and to various cellular processes (Supplementary Data Sets 1-4 and Table 2). For example, it was recently reported that two dendritic mRNAs, *PSD-95* and *CaMKIIa*, possessed 3'-UTR G4s with the ability to act as specific localization signals, targeting these RNAs to cortical neurites[17]. Moreover, the FMRP protein, already known to bind G4 structures, has been suggested to act as one of the *trans*-acting factors in this phenomenon[37]. In the present study, the PG4 sequences found in the 3'-UTRs of the *LRP5* and *FXR1* mRNAs were demonstrated to fold into G4 structures *in vitro* in the presence of a physiological concentration of KCl. Once in their 3'-UTR's natural context, and cloned downstream of a luciferase reporter gene, both of these G4 structures were shown to increase gene expression by 2-fold (Fig. 2b and 3f). These increases were associated with a more efficient polyadenylation at sites located few nucleotides upstream of the G4 structures (Fig. 2 and 3).

In metazoans, a polyadenylation unit is composed of various RNA elements located near its cleavage site[21]. Among the most common downstream elements are the U/GU-rich and the G-rich auxiliary elements. In light of the results presented here, the 3'-UTR G4 structures most likely act as downstream auxiliary elements that enhance the productivity of alternative polyadenylation sites. Two prevalent models have been proposed for the functionality of such auxiliary elements[38]. First, these elements could promote processing efficiency by

maintaining the core PAS in an unstructured form, thus enabling a better assembly of the general polyadenylation factors. In this regard, the extreme stability of RNA G4 structures may be a favorable characteristic. Additionally, in-line probing results typically showed that the regions flanking the G4 structure become both more flexible and single-stranded upon its formation (Fig. 1a,d and Supplementary Fig. 1a,d)[8]. Second, the auxiliary elements could interact with specific proteins, which would in turn stimulates the assembly of the general polyadenylation factors on the pre-mRNA. For example, it has been reported that a G4 structure, located in 3' of the *p53* gene, was essential in maintaining the efficient 3'-end processing of the pre-mRNA under stress-induced DNA damage throughout the interaction with hnRNP H/F20. Undoubtedly, many characteristics of the G4 structure make it a suitable candidate to act as a polyadenylation auxiliary element.

Over one hundred mRNAs were shown to harbor putative alternative polyadenylation units composed of either an AAUAAA or an AUUAAA polyadenylation signal and a 3'-UTR PG4 (Supplementary Data Set 5). This is most likely an under-estimation as there are many other known variant polyadenylation signals in mammalian cells[39,40] and the distance used here (100 nt) is minimal considering that G-rich regulatory elements located as far as 440 nt downstream of the core polyadenylation site of a mRNA have already been shown to be critical for efficient 3'-end processing[41]. In addition to these facts, an enrichment of PG4 sequences located in the first 10% of the 3'-UTRs (i.e. near downstream stop codons) was observed, suggesting that the deletion of larger sequences is favored (Supplementary Fig. 2). Most likely only the « tip of the iceberg », in terms of G4 structures that may act as auxiliary alternative polyadenylation elements, has been revealed.



Chapitre 3, Figure S 2. Distribution of the 3'-UTR PG4 sequences.

All 3'-UTRs of the UTRef collection were fractionated into ten equal parts. All PG4 sequences found on the complementary strand were localized and their positions analyzed. The x-axis identifies each portion of the 3'-UTR and the y-axis the percentage (%) of PG4 sequences found in each portion.

The study of two different candidates permitted evaluation of the impact of the 3'-UTR G4 in two distinct contexts. In the case of the *LRP5* 3'-UTR, the polyadenylation unit containing the G4 structure was the only efficient one. The modulation of its efficiency by the G4 directly determined the level of mRNA produced and, consequently, the level of protein synthesized (Fig. 2b). The impact of the G4 promoting alternative polyadenylation was significantly different in the *FXR1* 3'-UTR environment, and provides a quick overview of how complex this mechanism can be. The *FXR1* 3'-UTR contains both an alternative and a canonical polyadenylation units resulting in the production of a short and a long isoform, respectively (Fig. 3). A tight coordination between both polyadenylation units was observed. The mRNA with a wt G4 structure favored the short isoform, while an mRNA with the G/A-mutated version accumulated more of the long isoform. The overall impact of the modification of this short/long isoform ratio was an increase in

the level of protein produced in the presence of the G4 structure. This observation is in accordance with the notion that an mRNA with a shorter 3'-UTR is usually both more stable and more actively translated than is one with a longer 3'-UTR[27]. Moreover, living cells use shortened 3'-UTRs to increase the expression of various genes during specific processes, such as proliferation and oncogene activation, without genetic alteration[26,27]. The better translational efficiency is a consequence of the loss of 3'-UTR repressive elements, mainly microRNA binding sites. In agreement with this, the loss of a shared microRNA seed region located in position 813 of the *FXR1* 3'-UTR appeared to be responsible for the better translational properties of the shorter isoform, this in a process in which miR-92b has a significant role (Fig. 4). With 3'-UTR mRNA shortening attracting a lot of attention recently[26,27], the G4 structures located in 3'-UTR may gain in popularity as an RNA motif to study for a better understanding of this phenomenon.

The characterization of the *FXR1* 3'-UTR also demonstrated that the amount of mRNA synthesized at the level of the downstream canonical polyadenylation site seems to be independent of the use of the alternative upstream site (Fig. 3a-e). On the basis of this observation, it is tempting to speculate that the 3'-UTR G4 sequence may act also as a transcriptional termination element; however, additional physical support is required in order to confirm this hypothesis. That said, it is supported by studies reporting that G4s that form in the nascent RNA transcript stimulate mitochondrial transcription termination[42], and that G-rich regions were shown to form an R-loop which can act as a transcriptional pause site important in transcriptional termination in mammalian cells[42,43].

In summary, this study demonstrates that G4 structures are abundant within 3'-UTRs and that these RNA motifs appear to have diverse contributions to mRNA processing events such as alternative polyadenylation. In fact, looking at the G4 structures of two independent 3'-UTRs revealed that their impacts are considerably more complex than initially believed. This is nicely illustrated by the demonstration that the 3'-UTR G4 structure of the *FXR1* mRNA stimulates alternative polyadenylation and, consequently, leads to 3'-UTR shortening which in turn impairs its microRNA regulation and, ultimately, gene expression. In brief, G4

structures emerge as important *cis*-acting elements present in 3'-UTRs with important impacts on both alternative polyadenylation and gene expression.

ACKNOWLEDGMENTS

We are grateful to the laboratory of Marie-Paule Teulade-Fichou at the Institut Curie (France) for providing the PhenDC3 ligand, to Dominque Levesque for technical assistance and François Bachand for critical discussions. This work was supported by the Canadian Institutes of Health Research (to J.P.P). J.D.B. was the recipient of the CIHR Frederick Banting and Charles Best Canada Graduate Scholarships Doctoral Awards. J.P.P holds the Canada Research Chair in Genomics and Catalytic RNA.

AUTHOR CONTRIBUTIONS

J.D.B. conceptualized the study, designed and performed the experiments. J.D.B and J-P.P analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

MATERIALS AND METHODS

The sequences of the oligonucleotides used in this study are shown in Supplementary Table 1.

Bioinformatics. The human 3'-UTR databases were derived from sequences taken from UTRdb (UTRfull release 1 and UTRef release 9)[45]. PG4 sequences were identified using the algorithm mentioned in the text and the program RNAMotif[46]. The results were exposed to various homemade Perl scripts and manually cured to obtain the PG4 databases in an Excel file format. When a 3'-UTR PG4 was located in a gene that generates more than one transcript with the same 3'-UTR, each transcript was considered individually and was counted as one more PG4 (Supplementary Data Sets 1-2). Gene ontology and disease association were performed using the complementary 3'-UTR PG4 results from UTRef and the Database for Annotation, Visualization and Integrated Discovery (DAVID) web-accessible program (Supplementary Data Sets 3-5)[47]. The database of putative alternative polyadenylation units containing PG4 elements was constructed using homemade Perl scripts (Supplementary Data Set 5).

RNA synthesis and labeling. RNAs for the *in vitro* experiments were synthesized by transcription using T7 RNA polymerase as described both previously[48] and in Supplementary Methods.

Circular dichroism spectroscopy and thermal denaturation. Detailed procedures are provided in the Supplementary Methods.

In-line probing. In-line probings were performed as described previously[8]. Trace amounts of 5'-end-labelled RNA (<1 nM) were heated at 70°C for 5 min and then slow-cooled to room temperature over 1 h in buffer containing 50 mM Tris-HCl (pH 7.5) and with either no monovalent salt, or with either 100 mM LiCl, NaCl or KCl in a final volume of 10 µL. Following the slow-cooling, the volume of each sample was adjusted to 20 µL such that the final concentrations were 50 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$ and either no salt or 100 mM of either LiCl, NaCl or KCl. The reactions were incubated for 40 h at room temperature, and then 20 µL of formamide loading buffer (95% formamide and 10 mM EDTA) were added to each

sample. For alkaline hydrolysis, 5'-end-labeled RNA was dissolved in 5 µL of water, 1 µL of 1 N NaOH was added and the reactions incubated for 1 min at room temperature prior to being quenched by the addition of 3 µL of 1 M Tris-HCl (pH 7.5). The RNA in each sample was then ethanol-precipitated and dissolved in formamide loading buffer. The RNase T1 ladder was prepared using 5'-end-labeled RNA dissolved in 10 µL of buffer containing 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$ and 100 mM LiCl. The reactions were incubated for 2 min at 37°C in the presence of 0.6 U of RNase T1 (Roche Diagnostic), and were then quenched by the addition of 20 µL of formamide loading buffer. All of resulting samples were fractionated on denaturing (8M urea) 10% polyacrylamide gels. The gels were subsequently dried, visualized by exposure to phosphorscreens (GE Healthcare) and the radioactivity quantified using the SAFA software as described previously[8,31].

Cell culture, plasmids construction and DNA transfection. Detailed procedures are provided in the Supplementary Methods.

Dual luciferase assays. Twenty-four hours after the transfection of HEK293T cells (see Supplementary Methods), a 10% fraction of the transfected cells was lysed in 150 µL of Passive Lysis Buffer and used to measure both the firefly (Fluc) and rennilia luciferase (Rluc) activities using the Dual-luciferase Reporter Assay kit according to the manufacturer's protocol in a test tube using a GloMax 20/20 luminometer (Promega). For each lysate, the Fluc value was divided by the Rluc value. The ratios obtained for the wild-type version were compared to those obtained with the G/A-mutant version of each candidate and/or constructions harboring specific mutations (e.g. AltPAS-mut and miRseed-mut). Both the mean values and the standard deviations were calculated from at least three independent experiments.

RNaseH/Northern blot hybridizations and ribonuclease protection assays. Total cellular RNA was extracted from the remaining 90% of the transfected HEK293T cells using the TriPure Isolation Reagent (Roche Applied Science) according to the manufacturer's protocol. The extracted RNA (20 µg) was snap cooled in water in the presence of 300 ng of both a *Fluc* specific DNA oligonucleotide (see Supplementary Table 1 for sequences) and oligo-$(dT)_{12-18}$ (Invitrogen), or of only the *Fluc* specific oligonucleotide in a volume of 10 µL. After the snap cooling, RNase H 10X reaction buffer, 1 U of RNase H enzyme (Ambion) and water were added so as to obtain final concentrations of 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 50 mM NaCl, 0.5 mM EDTA, 1 mM DTT and 25 µg/mL BSA in a total volume of 15 µL. The samples were then incubated at 37°C for 1 h, and the reactions stopped by the addition of 15 µL of iced-cold formamide loading dye. ^{32}P-radiolabeled ladder was synthesized by *in vitro* transcription from the plasmid pPD1 as described previously[49]. Both the RNA samples (30 µL) and the ladder were fractionated on 6% denaturing (8 M urea) PAGE gels. Northern blots were hybridized using ^{32}P-5'-end-labeled either Fluc or *7SL* RNA specific DNA probes (see Supplementary Table 1 for the sequences) for 18 h at 42°C. The membranes were washed, exposed to a phophorscreen (GE Healthcare) and analyzed using a Typhoon apparatus (GE Healthcare) for detection and quantification. Precise polyadenylation sites were determined by 3'-RACE experiments.

Ribonucleic Protection Assay (RPA) were performed using 10 µg of total RNA extract and the RPA III™ Kit (Ambion) as recommended by the manufacturer. *Fluc* and *Rluc* specific probes with 15 nt 5'- and 3'-overhangs were transcribed from PCR products (see Supplementary Table 1 for the primers' sequences) for both the pGL3 and pRL-TK plasmids (Promega) which contain the *Fluc* and *Rluc* genes, respectively.

Statistical analysis. Analysis of a single dataset was done with a one-sample *t*-test in order to examine whether or not the means differed from the hypothetical value of 1. Comparison analysis was performed using an unpaired two-tailed *t*-test

assuming that the two populations had the same variances. All calculations were performed using GraphPad Prism 5.0, and $P < 0.05$ was considered as being significant.

REFERENCES

1. Birney, E. et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
2. Huppert, J.L. Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. *Chem Soc Rev* **37**, 1375–1384 (2008).
3. Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K. & Neidle, S. Quadruplex DNA: sequence, topology and structure. *Nucleic Acids Res.* **34**, 5402–5415 (2006).
4. Neidle,S. and Balasubramanian,S. (2006) Quadruplex nucleic acids. *RSC Publishing*, Cambridge.
5. Todd, A.K., Johnston, M. & Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **33**, 2901–2907 (2005).
6. Wong, H.M., Stegle, O., Rodgers, S. & Huppert, J.L. A toolbox for predicting g-quadruplex formation and stability. *J Nucleic Acids* **2010**, doi:10.4061/2010/564946 (2010).
7. Huppert, J.L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.* **33**, 2908–2916 (2005).
8. Beaudoin, J.-D. & Perreault, J.-P. 5'-UTR G-quadruplex structures acting as translational repressors. *Nucleic Acids Res.* **38**, 7022–7036 (2010).
9. Huppert, J.L., Bugaut, A., Kumari, S. & Balasubramanian, S. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.* **36**, 6260–6268 (2008).
10. Mani, P., Yadav, V.K., Das, S.K. & Chowdhury, S. Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. *PLoS ONE* **4**, e4399 (2009).
11. Lipps, H.J. & Rhodes, D. G-quadruplex structures: in vivo evidence and function. *Trends Cell Biol.* **19**, 414–422 (2009).
12. Du, Z., Zhao, Y. & Li, N. Genome-wide colonization of gene regulatory elements by G4 DNA motifs. *Nucleic Acids Res.* **37**, 6784–6798 (2009).

13. Verma, A. et al. Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. *J. Med. Chem.* **51**, 5641–5649 (2008).

14. Bugaut, A. & Balasubramanian, S. 5'-UTR RNA G-quadruplexes: translation regulation and targeting. *Nucleic Acids Res.* doi:10.1093/nar/gks068 (2012).

15. Kumari, S., Bugaut, A., Huppert, J.L. & Balasubramanian, S. An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. *Nat. Chem. Biol.* **3**, 218–221 (2007).

16. Ji, X. et al. Research progress of RNA quadruplex. *Nucl Acid Ther* **21**, 185–200 (2011).

17. Subramanian, M. et al. G-quadruplex RNA structure as a signal for neurite mRNA targeting. *EMBO Rep.* **12**, 697–704 (2011).

18. Gomez, D. et al. Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. *Nucleic Acids Res.* **32**, 371–379 (2004).

19. Marcel, V. et al. G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. *Carcinogenesis* **32**, 271–278 (2011).

20. Decorsière, A., Cayrel, A., Vagner, S. & Millevoi, S. Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage. *Genes Dev.* **25**, 220–225 (2011).

21. Millevoi, S. & Vagner, S. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Res.* **38**, 2757–2774 (2010).

22. Colgan, D.F. & Manley, J.L. Mechanism and regulation of mRNA polyadenylation. *Genes Dev.* **11**, 2755–2766 (1997).

23. Proudfoot, N. Poly(A) signals. *Cell* **64**, 671–674 (1991).

24. Tian, B., Hu, J., Zhang, H. & Lutz, C.S. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res.* **33**, 201–212 (2005).

25. Moore, M.J. From birth to death: the complex lives of eukaryotic mRNAs. *Science* **309**, 1514–1518 (2005).

26. Sandberg, R., Neilson, J.R., Sarma, A., Sharp, P.A. & Burge, C.B. Proliferating Cells Express mRNAs with Shortened 3' Untranslated Regions and Fewer MicroRNA Target Sites. *Science* **320**, 1643–1647 (2008).

27. Mayr, C. & Bartel, D.P. Widespread Shortening of 3'UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. *Cell* **138**, 673–684 (2009).

28. Paramasivan, S., Rujan, I. & Bolton, P.H. Circular dichroism of quadruplex DNAs: applications to structure, cation effects and ligand binding. *Methods* **43**, 324–331 (2007).

29. Lane, A.N., Chaires, J.B., Gray, R.D. & Trent, J.O. Stability and kinetics of G-quadruplex structures. *Nucleic Acids Res.* **36**, 5482–5515 (2008).

30. Regulski, E.E. & Breaker, R.R. In-line probing analysis of riboswitches. *Methods Mol. Biol.* **419**, 53–67 (2008).

31. Laederach, A. et al. Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments. *Nat Protoc* **3**, 1395–1401 (2008).

32. Halder, K., Largy, E., Benzler, M., Teulade-Fichou, M.-P. & Hartig, J.S. Efficient suppression of gene expression by targeting 5'-UTR-based RNA quadruplexes with bisquinolinium compounds. *Chembiochem* **12**, 1663–1668 (2011).

33. De Cian, A., Delemos, E., Mergny, J.-L., Teulade-Fichou, M.-P. & Monchaud, D. Highly efficient G-quadruplex recognition by bisquinolinium compounds. *J. Am. Chem. Soc.* **129**, 1856–1857 (2007).

34. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).

35. Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.* **11**, R90 (2010).

36. Cheever, A., Blackwell, E. & Ceman, S. Fragile X protein family member FXR1P is regulated by microRNAs. *RNA* **16**, 1530–1539 (2010).

37. Phan, A.T. et al. Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nat Struct Mol Biol* **18**, 796–804 (2011).

38. Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L. & Hovorun, D.M. Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. *Nucleic Acids Res.* **31**, 1375–1386 (2003).

39. Nunes, N.M., Li, W., Tian, B. & Furger, A. A functional human Poly(A) site requires only a potent DSE and an A-rich upstream sequence. *EMBO J* **29**, 1523–1536 (2010).

40. Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M. & Gautheret, D. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**, 1001–1010 (2000).

41. Dalziel, M., Nunes, N.M. & Furger, A. Two G-rich regulatory elements located adjacent to and 440 nucleotides downstream of the core poly(A) site of the intronless melanocortin receptor 1 gene are critical for efficient 3' end processing. *Mol. Cell. Biol.* **27**, 1568–1580 (2007).

42. Wanrooij, P.H., Uhler, J.P., Simonsson, T., Falkenberg, M. & Gustafsson, C.M. G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16072–16077 (2010).

43. Skourti-Stathaki, K., Proudfoot, N.J. & Gromak, N. Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell* **42**, 794–805 (2011).

44. Bechara, E. et al. Fragile X related protein 1 isoforms differentially modulate the affinity of fragile X mental retardation protein for G-quartet RNA structure. *Nucleic Acids Res.* **35**, 299–306 (2007).

45. Mignone, F. et al. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res.* **33**, D141–6 (2005).

46. Macke, T.J. et al. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucleic Acids Res.* **29**, 4724–4735 (2001).

47. Huang, D.W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57 (2009).

48. Beaudoin, J.-D. & Perreault, J.-P. Potassium ions modulate a G-quadruplex-ribozyme's activity. *RNA* **14**, 1018–1025 (2008).

49. Beaudry, D., Busière, F., Lareau, F., Lessard, C. & Perreault, J.P. The RNA of both polarities of the peach latent mosaic viroid self-cleaves in vitro solely by single hammerhead structures. *Nucleic Acids Res.* **23**, 745–752 (1995).

SUPPLEMENTARY METHODS

RNA synthesis and labeling

All of the PG4 versions used for the *in vitro* experiments were synthesized by *in vitro* transcription using T7 RNA polymerase as described previously[1]. Briefly, two overlapping oligonucleotides (2 µM each) were annealed and double stranded DNA was obtained by filling in the gaps using purified *Pfu* DNA polymerase in the presence of 5% dimethyl sulfoxide (DMSO). The duplex DNA containing the T7 RNA promoter sequence followed by the PG4 sequence was then ethanol-precipitated. After dissolution of the polymerase chain reaction (PCR) product in ultrapure water, run-off transcriptions were performed in a final volume of 100 µL using purified T7 RNA polymerase (10 µg) in the presence of RNase OUT (20 U, Invitrogen), pyrophosphatase (0.01 U, Roche Diagnostics) and 5 mM NTP in a buffer containing 80 mM HEPES-KOH, pH 7.5, 24 mM $MgCl_2$, 40 mM DTT and 2 mM spermidine. The reactions were incubated for 2 h at 37°C followed by a DNase RQ1 (Promega) treatment at 37°C for 20 min. The RNA was then purified by phenol:chloroform extraction followed by ethanol precipitation. RNA products were fractionated by denaturing (8 M urea) 10% polyacrylamide gel electrophoresis (PAGE; 19:1 ratio of acrylamide to bisacrylamide) using 45 mM Tris-borate pH 7.5 and 1 mM EDTA solution as running buffer. The RNAs were detected by UV shadowing, and those corresponding to the proper sizes of the PG4s were excised from the gel and the transcripts eluted overnight at room temperature in a buffer containing 1 mM EDTA, 0.1% SDS and 0.5 M ammonium acetate. The PG4s were then ethanol-precipitated, dried, dissolved in water and analyzed by spectrometry at 260 nm to determine their concentration.

In order to produce 5'-end-labeled RNA molecules, 50 pmol of purified transcripts were dephosphorylated at 37°C during 1 h in the presence of 1 U of antartic phosphatase (New England BioLabs) in a final volume of 10 µL containing 50 mM Bis-Propane (pH 6.0), 1 mM $MgCl_2$, 0.1 mM $ZnCl_2$ and RNase OUT (20 U, Invitrogen). The enzyme was inactivated by a 5 min incubation at 65°C.

Dephosphorylated RNAs (5 pmol) were 5'-end-radiolabeled using 3 U of T4 polynucleotide kinase (Promega) for 1 h at 37°C in the presence of 3.2 pmol of [γ-^{32}P]ATP (6000 Ci/mmol; New England Nuclear). The reactions were stopped by the addition of formamide dye buffer (95% formamide, 10 mM EDTA, 0.025% bromophenol blue and 0.025 xylene cyanol), and the RNA molecules were then purified by 10% polyacrylamide 8 M urea gel electrophoresis. The bands containing the 5'-end-labeled RNAs were detected by autoradiography and those corresponding to the correct sizes were excised and recovered as described above.

Circular dichroism spectroscopy and thermal denaturation

All circular dichroism (CD) experiments were performed using 4 μM of the appropriate RNA transcript dissolved in 50 mM Tris-HCl (pH 7.5) either in the absence of monovalent salt, or in the presence of 100 mM of LiCl, NaCl or KCl. Prior to all CD measurements, all samples were heated in a water bath at 70°C for 5 min and then slow-cooled to room temperature over 1 h. CD spectroscopy experiments were performed using a Jasco J-810 spectropolarimeter equipped with a Jasco Peltier temperature controller in a 1-mL quartz cell with a pathlength of 1 mm. The CD scans were recorded ranging from 220 to 320 nm at 25°C at a rate of 50 nm min^{-1} and with a 2-s response time, 0.1-nm pitch and 1-nm bandwidth. The means of at least three wavelength scans were collected. Substraction of the buffer was not required since control experiments in the absence of RNA showed negligible curves. CD melting curves were recorded by heating the samples from 25°C to 90°C at a controlled rate of 1°C min^{-1} and monitoring a 264-nm CD peak every 0.2 min. Melting temperature (T_m) values were calculated using "fraction folded" (θ) *versus* temperature plots[2].

Cell culture

HEK 293T cells (human embryonic kidney) were cultured in T-75 flask (Sarstedt) in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine

serum (FBS), 1 mM sodium pyruvate and an antibiotic-antimycotic drug mixture (all purchased from Wisent) at 37°C in a 5% CO2 controlled atmosphere in a humidified incubator.

Plasmids construction

The Fluc-*LRP5* and Fluc-*FXR1* constructions were built based on 3'-UTR sequences from the NCBI database (i.e. NM_002335 and NM_005087, respectively; UTRdb Locus 3HSAA093364 (UTRfull) and 3HSAR019368 (UTRef), respectively). The full-length 3'-UTR of *LRP5* was reconstituted *in vitro* by the filling in of multiple overlapping oligonucleotides and various PCR steps. For the *FXR1* constructions, a plasmid containing the *FXR1* 3'-UTR was purchased from plasmID DF/HCC DNA Resource Core (HsCD00334849) and was used as template for PCR amplification with the proper forward and reverse oligonucleotides (see Table S1). 3'-UTRs harboring either the wild-type or the G/A-mutant G4 versions were synthesized for each candidate. Site directed mutagenesis was used to build constructions with alternative PAS mutations (*LRP5* AAUAAA to ACUAAC and *FXR1* AUUAAA to ACUAAC) and *FXR1* miRseed mutationt (UGUGCAAU to CCUGUUAG). The list of oligonucleotides used for each candidate is shown in Table S1. The reconstituted 3'-UTRs were double digested with Xba I and BamH I for the *LRP5* constructions, and Xba I and Sal I for the *FXR1* constructions. Digestion products were inserted into the pGL3 control vector plamid (Promega) previously digested with the same enzymes. DNA sequencing of each construction confirmed the insertion of the correct sequence.

DNA transfection

Typically, HEK 293T cells (6 x 105) were seeded in 6-well plates. The cells were co-transfected 24 h later with both the specific pGL3-control plasmid (firefly luciferase, *Fluc*) and the pRL-TK plasmid (renillia luciferase, *Rluc*) (Promega) using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. After an additionnal 24 h, 10% of the cells were used to measure the Rluc and Fluc

activities using the Dual-luciferase Reporter Assay kit (Promega). Total cellular RNA was extracted from the remaining cells (90%) using TriPure Isolation Reagent (Roche Applied Science) according to the manufacturer's protocol. Harvested total RNA was used for RNaseH/Northern blot hybridization and RPA experiments.

In order to test the impact of the G4 specific ligand (PhenDC3), HEK 293T cells (6 x 10^4) were seeded in 48-well plates and co-transfected 24 h later as described above. Various concentrations of PhenDC3 were then added to the cells four hours after the transfection. All of the cells were collected 24 h latter and subjected to Dual-luciferase assays.

In order to investigate the impacts of an inhibitor specific for miR-92b and of an irrelevant inhibitor control, HEK 293T cells (6 x 10^4) were seeded in 48-well plates. 24 h later, the cells were initially transfected with 100 nM (final concentration) of either the specific miR-92b inhibitor or the irrelevant inhibitor control using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. Two hours post-transfection, the cells were then co-transfected with the specific Fluc and Rluc constructions using Lipofectamine 2000 as described above. All of the cells were collected 24 h post transfection and subjected to Dual-luciferase assays.

Chapitre 3, Table S 1. Oligonucleotides used in this study.

For each candidate, the table provides both the names and the sequences of the oligonucleotides required for the synthesis of the different PG4 versions, as well as for the construction of the different 3'-UTR versions. The oligonucleotides used for each technique are also provided.

Candidate	Name	5'- Sequence -3'
LRP5	3UTR-1	TCAGTCTAGACCTCGGCCGGGCCACTCTGGCTTCTCTGTGCCCCTGTAAATAGTTTTAAATATGAACAAAGAAAAAAATATATTTGATTTAAAAAAT
	3UTR-2-G4Wt	GTACAAAGTTCTCCCAGCCCTGCCCACCCCATCACAGTTCACATTTCTCATGTTTTTAAAATCCCAATTATATTTATTTTTTAAATAAAATATATTT
	3UTR-2-G4Mut	GTACAAAGTTCTCTCAGCTCTGCTCACTCCATCACAGTTCACATTTCTCATGTTTTTAAAATCCCAATTATATTTATTTTTTAAATAAAATATATTT
	3UTR-3	TCGAGGATCCCTGTTTTACAAAATTAAGTTTATAAATATTTCTCCACTGTACAAAGTTCTC
	AltPASmut-Forward	GATTTAAAAACTAACTATAATTGGGATTTTAAAAACATGAGAAATGTG
	AltPASmut-Reverse	CCCAATTATAGTTAGTTTTTAAATCATAAAATATATTTTTTTCTTTGTTC
	PG4 wt	CTGTACAAAGTTCTCCCAGCCCTGCCCACCCCATCACAGTTCACATTTCCTATAGTGAGTCGTATTA
	PG4 G/Amut	CTGTACAAAGTTCTCTCAGCTCTGCTCACTCCATCACAGTTCACATTTCCTATAGTGAGTCGTATTA
FXR1	3UTR-Forward	GATCTCAGTCTAGAACTGAAGAAGTTCCTAGTTTACAG
	3UTR-Reverse	GATCTCAGGTCGACTTCAGCAAATGGAAGATCAAAG
	G4mut-Forward	CAAAATATGGCAGAGATGGAGAGTGGTGAGTGAGGTGAGATCCCTAACGTAATATTC
	G4mut-Reverse	GAATATTACGTTAGGGATCTCACCTCACTCACCACTCTCCATCTCTGCCATATTTTG
	AltPASmut-Forward	CCTTGAGAATAGTATATGTAACACTAACAAAAAGTTGCTGGCTATAGGAAATG
	AltPASmut-Reverse	CATTTCCTATAGCCAGCAACTTTTTGTTAGTGTTACATATACTATTCTCAAGG
	miRseed-mut-Reverse	GATCACGTCGACTTCAGCAAATGGAAGATCAAAGTTTATTTACTACCTGCATACAAAAACATCTAACAGGTCAAACAACCGAAAAAA
	PG4 wt	TATTACGTTAGGGATCCCACCCCACCCACCACCCCCCATCTCTGCCATATTTTGCCCTATAGTGAGTCGTATTA
	PG4 G/Amut	TATTACGTTAGGGATCTCACCTCACTCACCACTCTCCATCTCTGCCATATTTTGCCCTATAGTGAGTCGTATTA
RNaseH/ Northern	Fluc-RNaseH-cds-oligo	CCACAAACACAACTCCTCC
	Fluc-cds-Northern-probe	GGATCTCTCTGATTTTTCTTGCGTCGAG
	h7SL RNA-Northern-probe	ACCCGATCGGCATAGCGCACTACAGC
RPA	Fluc-+15-Forward	TGGCCTCCGTCTGCGGGCCCGGCGCCATTCTATCCGCTGG
	Fluc-+15-Reverse	GATCTAATACGACTCACTATAGGGAGGCCACCGACAGCGGCCAACCGAACGGACATTTCG
	Rluc-+15-Forward	TGGCCTCCGTCGTCCGGTATGGGCAAATCAGGCAAATCTGG
	Rluc-+15-Reverse	GATCTAATACGACTCACTATAGGGAGGCCACCGACGACCCCCATGATTCAATCACATCTACTAC
3'RACE	Fluc-cds-Forward	GGTCTTACCGGAAAACTCGACG
	FXR1-3'UTR-Forward	GACTGTTCTACCTTGAGGC

Tous les Datasets peuvent être retrouvés dans le fichier .zip

Supplementary Data Set 1 | PG4 database obtained from the UTRef collection.

Supplementary Data Set 2 | PG4 database obtained from the UTRfull collection.

Supplementary Data Set 3 | Gene ontology analysis results of the genes containing a 3'-UTR PG4s located on the complementary strand from the UTRef collection.

Supplementary Data Set 4 | OMIM database analysis for 3'-UTR PG4s located on the complementary strand from the UTRef collection related to various diseases.

Supplementary Data Set 5 | Analysis of PG4s located inside putative alternative polyadenylation units.

SUPPLEMENTARY REFERENCES

1. Beaudoin, J.-D. & Perreault, J.-P. Potassium ions modulate a G-quadruplex-ribozyme's activity. *RNA* **14**, 1018–1025 (2008).
2. Mergny, J.-L. & Lacroix, L. UV Melting of G-Quadruplexes. *Curr Protoc Nucleic Acid Chem Chapter 17*, Unit 17.1 (2009).

DISCUSSION

La majeure partie des résultats générés durant mon doctorat a été présentée et discutée dans les manuscrits du chapitre 1 à 3. Dans la discussion qui suit, je discuterai des résultats les plus importants, des questions qu'ils soulèvent et des perspectives auxquelles ils pavent la voie. Je m'appuierai majoritairement sur des observations personnelles, sur la littérature, ainsi que sur des résultats non publiés.

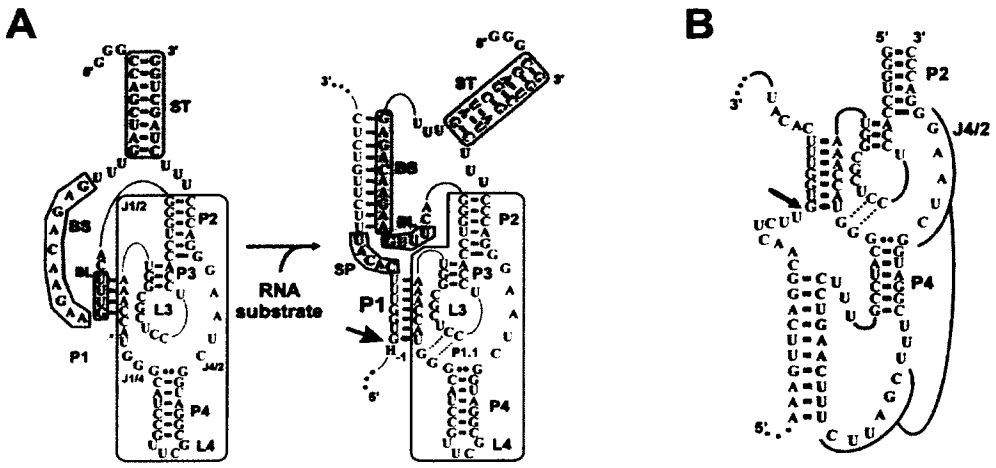
1. Mécanisme d'action du G-quartzyme

La chimère G-quadruplexe-ribozyme antigénomique du virus de l'hépatite D (VHD), que représente le G-quartzyme (GQRz), et son mécanisme d'action unique nous informe beaucoup sur l'efficacité de la structure G-quadruplexe à réguler l'activité de l'ARN. L'impact d'un G-quadruplexe sur l'activité du ribozyme VHD à catalyser le clivage d'un substrat d'ARN fût la première activité liée à l'ARN étudiée dans mes travaux.

1.1 Le processus d'inactivation

Premièrement, le GQRz nous rappelle qu'une structure G-quadruplexe est composée de différents niveaux de complexité, soient: primaire, secondaire et tertiaire. Chacun de ces niveaux semble pouvoir affecter de façon très différente un processus donné. Concernant le GQRz et son niveau de structure primaire, c'est-à-dire sa séquence nucléotidique, on peut remarquer qu'elle est à la base même du mécanisme d'inhibition de l'activité du ribozyme. En effet, la séquence du brin 5' de l'aptamère (G-quadruplexe) est capable de venir s'apparier à la séquence de la région P3 et L3 du ribozyme VHD, menant à l'inactivation de l'activité catalytique. Ceci entraine la formation d'une structure secondaire inhibitrice reposant sur des paires de bases Watson-Crick. Cette interaction est supportée par des résultats de cartographie, de cinétique enzymatique et d'analyse de mutants (Chapitre 1).

A priori, il peut sembler étonnant que deux motifs d'ARN aussi stable puissent s'affecter l'un et l'autre sans qu'aucun des deux ne soit replié correctement. La raison principale réside dans les conditions expérimentales. Effectivement, cette conformation inhibitrice est observée uniquement dans des conditions où la structure G-quadruplexe ne peut pas se former (c.-à-d. en absence de K^+). Dans un tel contexte, la séquence du G-quadruplexe est libre de former d'autres structures secondaires basées sur des paires de bases Watson-Crick. Le ribozyme VHD de son côté est tout à fait capable de maintenir un bon repliement et une bonne activité dans des conditions aussi extrême que 80°C, 5 molaire urée ou 18 molaire formamide (Doherty and Doudna, 2000; Shih and Been, 2002), mais il possède également quelques faiblesses. Afin d'être pleinement actif, le ribozyme VHD doit suivre un chemin de repliement bien précis et spécifique (voir Annexes 2.1 et 2.2). Les changements de conformations observés pendant ce cheminement impliquent et affectent majoritairement les nucléotides présents dans le cœur catalytique du ribozyme (c.-à-d. les régions P1, P1.1, P3, L3 et J4/2 (Chapitre 1, Figure 1)). De plus, la stabilité conférée par les tiges P2 et P4, situées en dehors du cœur catalytique, joue un rôle important dans la poursuite de ce chemin de repliement, particulièrement dans les conditions extrêmes mentionnées ci-dessus. Une étude d'ingénierie moléculaire, précédent celle du GQRz, avait démontré qu'il était possible d'inhiber l'activité du ribozyme VHD en ajoutant une région, appelée "bloqueur" (BL), capable de venir s'apparier à une partie de la région P1 (Discussion, Figure 1A)(Bergeron and Perreault, 2005). L'ajout d'une région bloqueur, mais également d'un biosenseur (BS) et d'un stabilisateur (ST), au ribozyme VHD antigénomique mena au développement d'une nouvelle classe de ribozyme, le ribozyme VHD-SOFA (SOFA pour "*specific on/off adapter*"). Elle démontra pour la première fois qu'il était possible d'interférer avec l'activité enzymatique du ribozyme VHD par l'utilisation d'une séquence d'ARN capable de venir s'apparier au cœur catalytique de celui-ci. Cette inhibition est levée en présence d'un substrat spécifique, reconnu par la partie biosenseur et la



Discussion, Figure 1. Différentes versions du ribozyme VHD-SOFA.

(**A**) et (**B**) Structure primaire et secondaire des ribozymes VHD-SOFA (**A**) et VHD-SOFA-*down* (**B**). Pour le ribozyme VHD-SOFA, les conformations "*OFF*" (à gauche) et "*ON*" (à droite) sont représentées. La section grise indique le module SOFA et la section blanche le ribozyme VHD. Les différents segments du module SOFA sont encadrés: le biosenseur (BS), le bloqueur (BL) et le stabilisateur (ST). L'ajout du substrat permet la transition de la conformation "*OFF*" à "*ON*". (**B**) Uniquement la conformation "*ON*" du ribozyme VHD-SOFA-*down* est représentée. La partie ombragée indique le module SOFA-*down*. L'interaction inhibitrice est démontrée entre le module SOFA-*down* et la portion J4/2 du ribozyme VHD. Figure adaptée de (Bergeron et al., 2005).

tige P1 du ribozyme VHD-SOFA. Le module SOFA augmente alors la fidélité du ribozyme en augmentant le nombre de nucléotides impliqués dans la reconnaissance du substrat et en maintenant une conformation inactive, grâce au bloqueur, en absence de celui-ci. Ici, le substrat agit comme effecteur positif sur l'activité du ribozyme. Suivant la même logique, une autre version du ribozyme VHD-SOFA a été développée, le ribozyme VHD-SOFA-*down* (Discussion, Figure 1B)(Bergeron et al., 2005). Ce dernier agissant de la même façon que la version précédente, à l'exception que la région du bloqueur s'apparie à la région J4/2 du ribozyme au lieu de la région P1. Ces deux versions de ribozymes SOFA ont une chose en commun, elles reposent initialement sur des conformations inactives, où le bloqueur s'apparie à une région importante du cœur catalytique du ribozyme, inhibant l'activité enzymatique de ce dernier (Discussion, Figure 1). En résumé,

l'appariement d'une séquence d'ARN au cœur catalytique du ribozyme VHD semble être un excellent moyen de l'inactiver. Dans le cas du GQRz, il fût le premier exemple qu'une interaction au niveau de la tige P3 et L3 est également très efficace afin d'observer le même phénomène. Une étude ultérieure de notre laboratoire, portant sur l'optimisation du ribozyme VHD-SOFA comme outil moléculaire, démontra que la présence d'un biosenseur possédant une séquence capable de venir former aussi peu que six paires de bases avec les régions P3 et L3 du ribozyme réduisait grandement son activité (Lévesque et al., 2011). La possibilité d'une interaction entre ces deux régions a été identifiée pour constituer un critère important à vérifier lors du développement de ribozymes VHD-SOFA efficaces pouvant agir comme outils moléculaires. Dans le cas du GQRz, en plus d'observer une structure inhibitrice via l'interaction avec le cœur catalytique, l'ajout de l'aptamère dans la région P2 du ribozyme semble déstabiliser la tige P2. Cette déstabilisation est facilement observable avec la cartographie à la RNase T1 (Chapitre 1, Figure 4A,B). On peut voir avec cette dernière, que les guanines du brin 3' de l'aptamère, particulièrement les guanines 68 et 70, sont plus accessibles dans la conformation inactive et deviennent moins accessibles en présence de K^+. Ceci suggère fortement que la tige P2 n'est pas formée en absence de K^+ ajoutant un autre élément inhibiteur de taille. En conclusion, l'interaction du brin 5' de l'aptamère avec le cœur catalytique du ribozyme et la déstabilisation de la tige P2 sont deux phénomènes cruciaux responsables de la perte d'activité du ribozyme en absence de contre-ion et sont principalement régis par la structure primaire du GQRz.

1.2 Le processus d'activation

La compréhension du mécanisme d'inactivation du GQRz soulève déjà plusieurs pistes intéressantes sur la capacité d'une structure G-quadruplexe à affecter une activité de l'ARN. Toutefois, afin d'être identifiée comme régulateur de l'ARN, il faut que la structure G-quadruplexe soit directement impliquée dans le passage d'un état d'activité X à un état d'activité Y. Concernant le GQRz, l'ajout ou la présence

de K^+ est l'élément déclencheur permettant le passage d'une très faible à une très forte activité catalytique de clivage du GQRz (Chapitre 1, Figure 1B,C). La présence du contre-ion K^+ permet très clairement la formation d'une structure G-quadruplexe au niveau de la partie aptamère du GQRz. Cette formation est supportée par les cartographies à la RNase T1, au DMS, l'analyse des différents mutants ainsi que l'impact du potassium et d'un ligand de G-quadruplexe ($TMPyP_4$) sur l'activité catalytique du GQRz. Ces résultats démontrent clairement que la formation d'un G-quadruplex impliquant plusieurs guanines de la partie aptamère empêche et/ou défait la structure inhibitrice produite par l'association du brin 5' de l'aptamère à la région P3 et L3 du ribozyme, permettant un bon repliement du cœur catalytique. En plus, le G-quadruplexe aide à la stabilisation de la tige P2 comme on peut le voir sur le gel de cartographie à la RNase T1 (Chapitre 1, Figure 4A,B). Ces deux aspects, très bénéfiques à l'activité du ribozyme, permettent au GQRz de gagner énormément en efficacité en présence de K^+. Si la conformation inhibitrice reposait majoritairement sur la structure primaire du G-quadruplexe (c.-à-d. sa séquence), la conformation active repose surtout sur sa structure secondaire et tertiaire, où la formation du G-quadruplexe a des répercussions indéniables et positives sur le motif d'ARN ribozyme VHD et son activité.

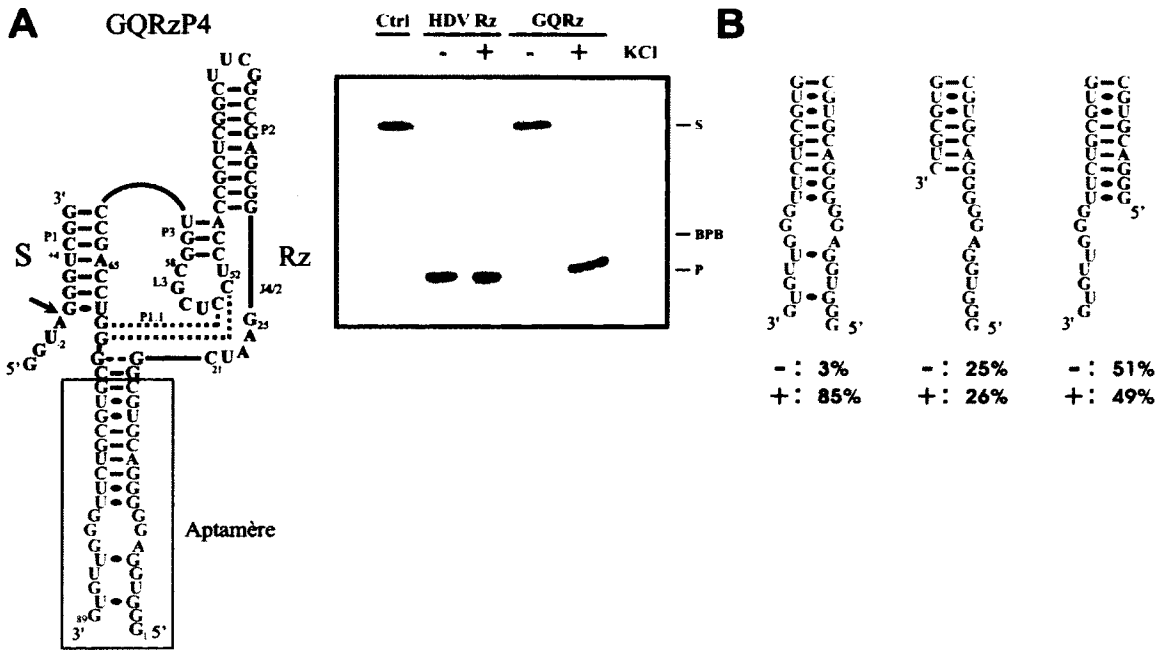
Une expérience clé concernant la modulation de la conformation inactive vers la conformation active est présentée à la Chapitre 1, Figure 2. Elle rapporte une information très importante. En effet, la totalité des autres expériences a été réalisée avec une étape de dénaturation et renaturation, où le GQRz, en présence du substrat d'ARN et dans une condition donnée (présence ou absence de K^+), est chauffé à une température de 70°C pendant deux minutes et graduellement refroidi à une température de 25°C sur une période d'une heure. Cette étape permet de faciliter le repliement du G-quadruplexe en défaisant les structures d'ARN et en laissant le temps aux structures plus stables de se former tranquillement. On peut s'imaginer que dans une condition favorable à la formation du G-quadruplexe, le GQRz adopte simplement une conformation active au détriment de l'inactive

puisqu'elle devient plus favorable thermodynamiquement. Toutefois, pour l'expérience en question, aucune dénaturation suivie d'une renaturation n'a été effectuée. Dans ce contexte, le GQRz se replie initialement dans une condition non favorable à la formation de G-quadruplexe résultant en une majorité de conformations inactives. Du Mg^{2+}, nécessaire à l'activité du ribozyme, est ajouté au temps zéro. Comme attendu, le ribozyme VHD de type sauvage clive très rapidement son substrat alors que le GQRz est principalement inactif (Chapitre 1, Figure 2). Cependant, après l'ajout de K^+ après 30 minutes, le GQRz gagne graduellement en activité pour finalement rejoindre un niveau de coupure similaire au ribozyme VHD de type sauvage. Tout ceci se produit à la même température, soit 37°C. Ces résultats démontrent que même sans dénaturation et renaturation, l'ajout de K^+ permet la transition d'une conformation inactive vers une conformation active. Dans le contexte du GQRz les forces thermodynamiques favorisant la formation du G-quadruplexe semblent être suffisamment puissantes pour défaire la structure inhibitrice, basée sur plusieurs paires de bases Watson-Crick, et permettre la formation du G-quadruplexe avec plusieurs de ces mêmes nucléotides, préalablement séquestrés. Cette possibilité de faire la transition d'une structure X vers une structure G-quadruplexe à une température de 37°C est extrêmement importante et intéressante afin d'imaginer un processus cellulaire régulé de cette façon. Cette caractéristique de la structure G-quadruplexe, dans le contexte du GQRz, commence à nous donner certaines pistes concernant les différents mécanismes potentiellement utilisés par un G-quadruplexe pour réguler un processus cellulaire axé sur l'ARN.

En conclusion, le mécanisme d'action du GQRz est un exemple très intéressant où une structure possédant une caractéristique particulière, ici une formation K^+ dépendante, est capable de la transférer à un autre motif d'ARN et de réguler son activité via ce caractère distinctif, ici le ribozyme et son activité de coupure. Ceci est une propriété que tout élément de régulation post-transcriptionnelle se doit de posséder et la structure G-quadruplexe semble parfaitement apte à le faire.

1.3 Flexibilité du mécanisme d'action du GQRz

Le GQRz tel que présenté au chapitre 1 de cette thèse propose un mécanisme d'action unique et original. Cependant, la question se pose à savoir si les paramètres de ce mécanisme ne sont pas limités à cette séquence et plutôt restreints? Il est possible que la structure G-quadruplexe étudiée réagisse de cette façon uniquement dans le contexte où elle est insérée dans la tige P2 du ribozyme VHD et nulle part ailleurs. Dans le but de mettre à l'épreuve la flexibilité du mécanisme d'action menant à la régulation de l'activité catalytique du ribozyme VHD par la formation d'une structure G-quadruplexe, une nouvelle version du GQRz a été développée, où l'aptamère a été fusionné au ribozyme via la tige P4 (résultats non publiés; Discussion, Figure 2). Ce nouveau GQRz possède également une activité K^+ dépendante. Les pourcentages de coupure en absence et en présence de K^+ sont très similaires pour chacune des deux versions de GQRz (Chapitre 1, Figure 1B et Discussion, Figure 2). L'analyse de mutants du GQRz version P4 (GQRzP4) démontre que, pour ce nouveau ribozyme, l'inhibition conférée par le brin 5' de l'aptamère est moins importante que celle observée pour la version P2. En effet, la délétion de ce brin n'a pas permis de retrouver la totalité de l'activité catalytique, comme c'était le cas pour la version P2 (comparer Chapitre 1, Figure 1C et Discussion, Figure 2B). Toutefois, l'activation dépendante du K^+ nécessite toujours la présence des deux brins de l'aptamère (et des guanines qu'ils contiennent) suggérant fortement que la formation du G-quadruplexe, similaire à celle observée pour la version P2, est responsable du gain d'activité (Discussion, Figure 2B). Dans le cas du GQRzP4, la déstabilisation de la tige P4 causée par l'insertion de l'aptamère semble être le facteur principal de la perte d'activité du ribozyme. En changeant le contexte du GQRz, il est possible de garder le même phénotype d'activité catalytique dépendante du K^+ et de modifier l'importance des différents facteurs impliqués dans le mécanisme d'action. Les résultats obtenus avec le GQRzP4 démontrent une certaine flexibilité de la structure G-quadruplexe à réguler une activité de l'ARN.



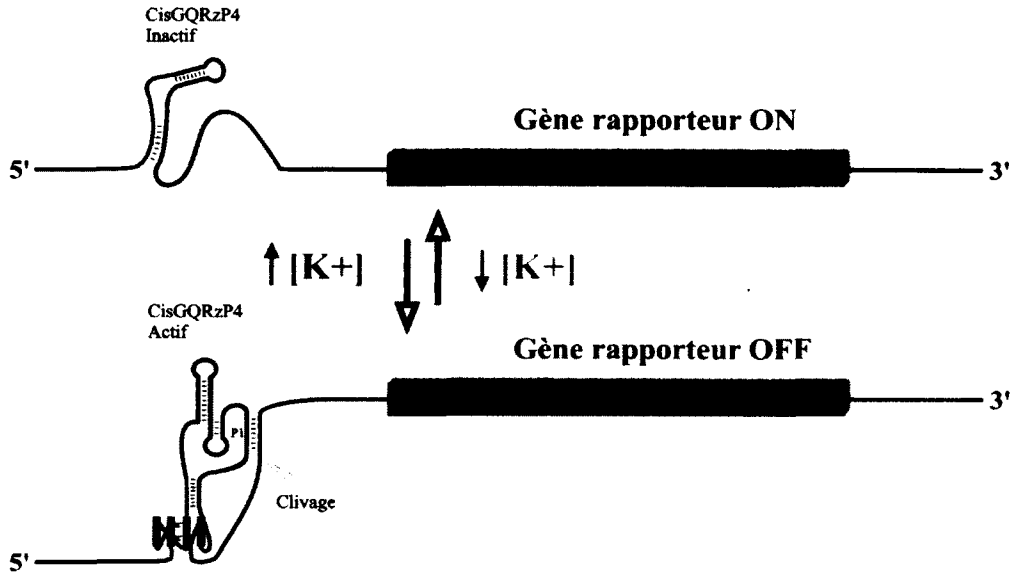
Discussion, Figure 2. Caractérisation du GQRzP4.

(**A**) Schéma du GQRz version P4. La section en orange représente la partie aptamère insérée dans la tige P4 du ribozyme VHD. La flèche indique le site de coupure. Un gel montrant des essais de coupure typiques du GQRzP4, avec 10 nM de ribozyme, 10 mM $MgCl_2$, des traces (<1 nM) de substrat radiomarqué en 5' et soit en absence ou en présence de 150 mM KCl, est présenté à droite. Les positions du substrat (S), produit (P) et bleu de bromophénol (BPB) sont indiquées à la droite du gel. Le contrôle négatif (Ctrl) est une incubation en absence de ribozyme. Le contrôle positif est le ribozyme VHD de type sauvage (HDV Rz) (**B**) Uniquement la portion correspondant à l'aptamère est représentée pour le GQRzP4 wt et deux mutants. L'activité, en pourcentage de coupure, après l'incubation de 10 nM de ribozyme soit en absence (-) ou en présence (+) de 150 mM de KCl pendant 1 h est présentée sous chaque molécule.

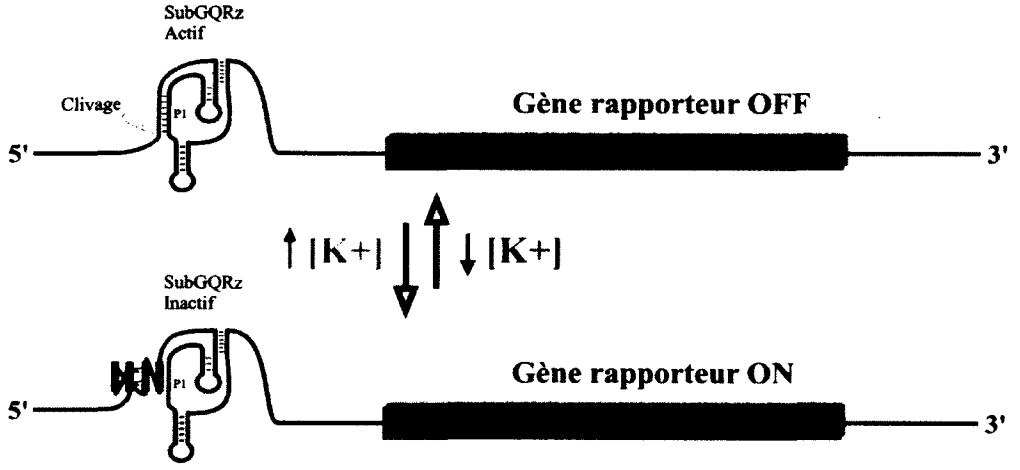
2. CisGQRz et modulation de l'expression génique

Le prochain défi auquel je me suis attaqué était de développer un système capable de moduler l'expression d'un gène basé sur le GQRz. Les deux versions discutées précédemment correspondent à des GQRz en *trans*, c'est-à-dire que le GQRz se trouve dans une molécule d'ARN et le substrat sur une autre. À l'inverse, on parle de ribozyme en *cis* lorsque les domaines ribozyme et substrat sont situés à l'intérieur de la même molécule d'ARN. Le clivage d'une molécule d'ARN par le

A **Mécanisme d'inactivation génique par le K^+**



B **Mécanisme d'activation génique par le K^+**



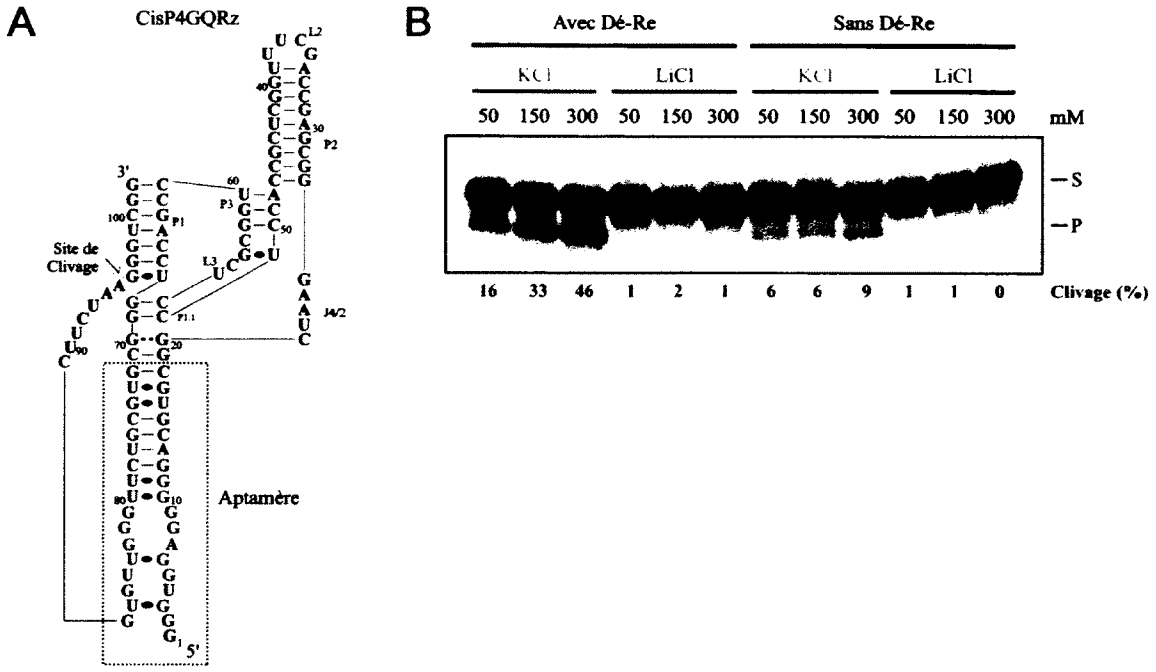
Discussion, Figure 3. L'utilisation de CisGQRz pour moduler l'expression génique.

(**A**) Schéma d'une construction utilisant le CisGQRzP4 afin de moduler l'expression d'un gène rapporteur. L'augmentation de l'activité du ribozyme, par la présence de K^+ ou de ligand, entrainerait une diminution de l'expression du gène rapporteur. (**B**) Schéma d'une construction utilisant le SubGQRz afin de moduler l'expression d'un gène rapporteur. La diminution de l'activité du ribozyme, par la présence de K^+ ou de ligand, entrainerait une augmentation de l'expression du gène rapporteur.

ribozyme produit et libère deux molécules d'ARN, une avec une extrémité 5'-hydroxyl et une autre avec une extrémité 2'-3'-monophosphate cyclique. Ces deux extrémités sont reconnues rapidement par les exonucléases cellulaires et l'ARN ainsi clivé est donc dégradé rapidement en cellule. J'ai développé différentes versions du GQRz en *cis*, dans le but de pouvoir les insérer dans les régions 5'- ou 3'-UTR des ARNm afin que leur activité de coupure module directement la stabilité et, du même coup, le niveau d'expression de ces messagers. L'objectif ultime serait de développer deux systèmes: un où la formation du G-quadruplexe augmente l'activité du GQRz, menant à une diminution de l'expression génique (Discussion, Figure 3A), et un autre où la formation du G-quadruplexe diminue l'activité du GQRz, permettant une augmentation de l'expression génique (Discussion, Figure 3B).

2.1 Développement de versions en cis du GQRz en trans.

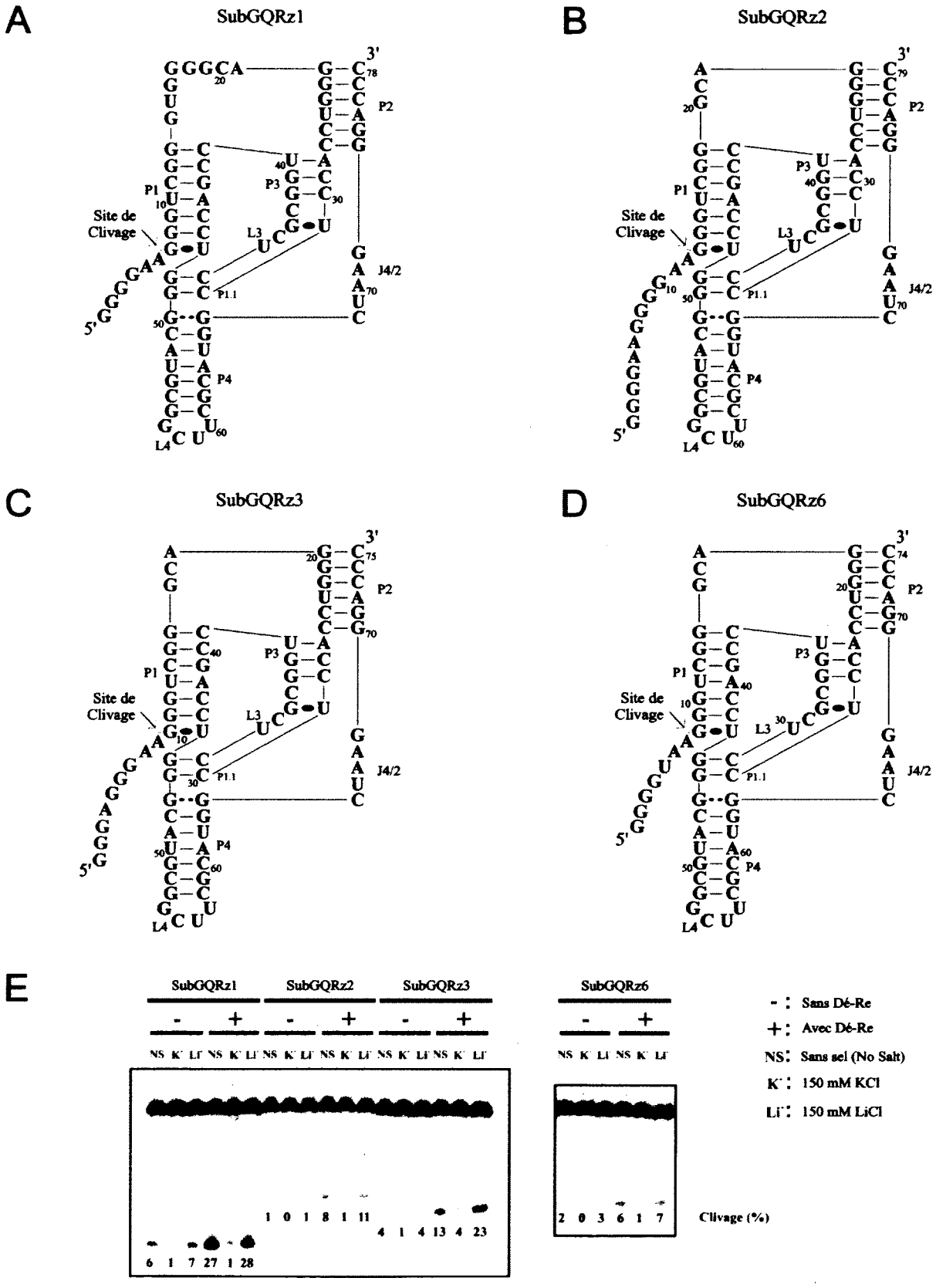
La première étape était de développer par ingénierie moléculaire deux types de GQRz en *cis* possédant ces caractéristiques (tous des résultats non publiés). Pour construire un GQRz en *cis* modulé positivement par la formation d'un G-quadruplexe, je suis parti du GQRzP4 en *trans* et j'ai fusionné l'extrémité 3' du GQRz à l'extrémité 5' du substrat. Ce nouveau ribozyme fût nommé le CisP4GQRz (Discussion, Figure 4). L'activité du CisP4GQRz a été calculée dans des conditions croissantes de K^+ ou de Li^+ (Discussion, Figure 4B). Une augmentation de l'activité catalytique accompagne l'augmentation de K^+ alors qu'aucune activité n'est observée en présence de Li^+. Le même phénomène est noté en présence ou en absence d'une étape de dénaturation-renaturation précédent l'essai de coupure, bien qu'une différence d'amplitude soit constatée (Discussion, Figure 4B). Il est difficile de prévoir laquelle des deux conditions (avec ou sans dénaturation-renaturation) est la plus représentative de ce qui se passera à l'intérieur d'une cellule. À ce moment, le repliement du GQRz s'effectuera bien différemment, c'est-à-dire de façon co-transcriptionnelle (Wong and Pan, 2009). Dans tous les cas, le



Discussion, Figure 4. Caractérisation du CisP4GQRz.

(**A**) Schéma du CisP4GQRz. L'encadré en pointillé identifie la partie aptamère du ribozyme. Le ribozyme (en *cis*) est relié au substrat via le brin 3' de l'aptamère et l'extrémité 5' du substrat. Les guanines en bleues sont celles potentiellement impliquées dans la structure G-quadruplexe. Le site de clivage est indiqué par la flèche. (**B**) Gel démontrant l'activité du CisP4GQRz radiomarqué en 5' en présence de différentes concentrations de KCl ou LiCl, soit 50, 150 ou 300 mM. Les résultats pour des ribozymes ayant traités par une étape de Dénaturation-Renaturation (Dé-Re) ou non, précédent l'essai de coupure, sont présentés. Les positions du substrat (S) et du produit (P) sont indiquées à la droite du gel et les pourcentages de coupure en dessous du gel.

phénotype observé est celui désiré car l'activité du GQRz augmente avec la formation du G-quadruplexe, dictée ici par la concentration en potassium. Le deuxième type de GQRz en *cis* à développer contraste complètement avec ce qu'on a vu jusqu'à maintenant, c'est pourquoi j'ai décidé d'adopter une stratégie complètement différente. L'idée était d'interférer avec l'étape initiale du chemin réactionnel du ribozyme, soit la reconnaissance du substrat. Le ribozyme utilise une région de sept nucléotides afin de reconnaître son substrat via des paires de bases Watson-Crick, formant la tige de reconnaissance P1. La séquence retrouvée


A
SubGQRz1
B
SubGQRz2
C
SubGQRz3
D
SubGQRz6
Site de Clivage
P1
P2
P3
P4
P1.1
L3
L4
J4/2
5'
3'
E
SubGQRz1
SubGQRz2
SubGQRz3
SubGQRz6
- : Sans Dé-Re
+ : Avec Dé-Re
NS : Sans sel (No Salt)
K+ : 150 mM KCl
Li+ : 150 mM LiCl
Clivage (%)

Discussion, Figure 5. Caractérisation des SubGQRz.

(**A-D**) Schéma de différentes versions de SubGQRz. Le ribozyme (en *cis*) est relié au substrat via l'extrémité 5' de la tige P2 et l'extrémité 3' du substrat. Les guanines en bleues sont celles potentiellement impliquées dans la structure G-quadruplexe. Le site de clivage est indiqué par la flèche. (**E**) Gel démontrant l'activité des différents SubGQRz radiomarqués en 5' soit en absence de sel (NS) ou en présence de 150 mM de KCl (K^+) ou LiCl (Li^+). Les résultats pour des ribozymes ayant subis une étape de Dénaturation-Renaturation (+) ou non (-) précédent l'essai de coupure sont présentés. Les pourcentages de coupure sont présentés en dessous du gel.

pour le ribozyme de type sauvage est CCGACCU (dans l'orientation 5' vers 3') alors que la séquence de la partie substrat de cette tige est GGGUCGG (dans l'orientation 5' vers 3')(Discussion, Figure 5A-D). J'ai donc décidé d'utiliser les deux séries de guanines présentes dans la partie substrat de la tige P1 et d'ajouter des guanines additionnelles en 5' et 3' de celles-ci. Ces additions permettent l'apparition de différentes compositions de quatre séries de guanines avec un fort potentiel à former une structure G-quadruplexe (Discussion, Figure 5A-D). Le ribozyme a besoin que la partie substrat de la tige P1 soit simple brin afin de la reconnaître de façon efficace, alors, si celle-ci forme une structure G-quadruplexe, le ribozyme sera incapable de lier son substrat et d'entrainer son clivage. Pour les quatre versions développées (appelées SubGQRz1,2,3 et 6), une activité de coupure est observée en absence de sel ou en présence de Li^+ alors qu'aucune n'est présente en présence de K^+ (Discussion, Figure 5E). Une fois de plus, le même phénomène est observé en présence ou en absence d'une étape de dénaturation-renaturation. En utilisant cette stratégie innovatrice, il a été possible de développer des GQRz possédant un niveau d'activité inversement proportionnel à la formation d'un G-quadruplexe. L'activité de ces GQRz en *cis* est toujours dépendante de la concentration en K^+, mais de façon inverse à celle qui a été démontrée précédemment. En résumé, il a été possible de développer deux types différents de GQRz en *cis* dont l'activité de chacun est dépendante du potassium bien que de façon diamétralement opposée. Cette possibilité ajoute encore davantage à la flexibilité que détient la structure G-quadruplexe à agir comme modulateur de l'activité de l'ARN dans différents contextes.

2.2 Utiliser la bactérie comme organisme modèle

La démonstration de ces nouveaux types de GQRz a été réalisée *in vitro*. La prochaine étape est de les utiliser afin de moduler l'expression génique *in cellulo* ou *in vivo*. Comme mentionné ci-dessus, les systèmes préconisés consisteraient à insérer chacun des différents GQRz dans le 5'-UTR d'un gène rapporteur (p. ex. la luciférase, la "green fluorescent protein" (GFP) ou la beta galactosidase) afin que l'activité des GQRz soit inversement proportionnelle à la quantité de protéines produites (Discussion, Figure 3). La caractérisation *in vitro* a été effectuée en faisant varier la concentration en K^+ afin de favoriser ou non la formation de la structure G-quadruplexe et, par le fait même, l'activité du GQRz. Cependant, il est très difficile de faire varier la concentration intracellulaire de K^+ des cellules de mammifères, qui se situe initialement entre 100 et 150 mM dépendamment du type cellulaire. Dans ce cas précis, les cellules de mammifères ne semblent pas constituer un bon organisme modèle. C'est pourquoi nous sommes présentement en collaboration avec le Dr. Wade Winkler, spécialiste de l'étude des riborégulateurs chez les bactéries, afin de mener plus loin cette partie du projet. Certaines bactéries sont capables de produire des spores lorsque les conditions environnementales deviennent hostiles à leur croissance cellulaire. Ce processus de sporulation est très complexe et utilise une machinerie cellulaire bien spécialisée pouvant mener à la formation d'un manteau protecteur de la spore contenant un mélange de plus de 30 polypeptides (Barák et al., 2005; Higgins and Dworkin, 2012). Ce véritable bouclier permet à la cellule végétative de subsister jusqu'à ce que les conditions environnementales redeviennent propices à la croissance cellulaire. Pendant la sporulation la concentration de plusieurs ions monovalents et divalents varie énormément. En fait, une des premières étapes de ce phénomène est l'accumulation massive de K^+ à l'intérieur de la cellule. Par exemple, un choc osmotique modéré chez *bacillus subtilis*, produit par la présence de 400 mM de NaCl dans le milieu de culture, entraîne une augmentation de la concentration intracellulaire de K^+ de son niveau basal, d'environ 315 mM, à 650 mM en seulement une heure (Kempf and Bremer, 1998). La présence de phénomènes nécessitant ou produisant une grande variation de la concentration

intracellulaire de K^+ chez la bactérie pourrait permettre la caractérisation des GQRz en *cis* comme régulateur de l'expression génique dans ces différentes conditions. Ils seraient en mesure de représenter un prototype de riborégulateur artificiel dépendant du potassium. Leur étude *in vivo* pourrait éventuellement mener à l'établissement de certains critères importants et, ultimement, mener à la découverte de riborégulateurs naturels utilisant le K^+ comme ligand.

2.3 Utiliser les GQRz en cis pour l'étude de ligands in cellulo

Comme il a été démontré au chapitre 1, le K^+ n'est pas le seul effecteur positif à l'activité du GQRz. La porphyrine ($TMPyP_4$) s'avère également efficace à augmenter l'activité du GQRz. En effet, en présence d'une concentration de K^+ similaire à celle retrouvée dans les cellules de mammifères (100-150 mM) la présence de 250 nM de $TMPyP_4$ mène à une augmentation de plus de deux fois de l'activité du GQRz (Chapitre 1, Figure 3C). À la suite de ces résultats, il est parfaitement plausible de penser que les systèmes d'expression génique présentés ci-dessus, incluant les différents types de GQRz en *cis*, pourraient être utilisés afin de tester l'efficacité de plusieurs ligands à lier un G-quadruplexe *in cellulo* (plus précisément en cellules de mammifères). Un engouement réel est apparu depuis déjà plusieurs années afin de développer des ligands spécifiques à la structure G-quadruplexe (Collie and Parkinson, 2011; Düchler, 2012). Étant donné les rôles variés des G-quadruplexes dans plusieurs processus cancéreux, bon nombre de ces ligands démontrent des propriétés anti-tumorales. À ma connaissance, bien qu'il existe différentes méthodes dans le but de cribler rapidement l'efficacité d'un grand nombre de ligands à lier spécifiquement une structure G-quadruplexe comparativement à un duplex *in vitro*, aujourd'hui, aucun système comparable ne permet de le faire *in cellulo* à grande échelle (Teulade-Fichou et al., 2007; Monchaud et al., 2008; Lacroix et al., 2011). Un tel système *in cellulo* aurait non seulement l'avantage d'étudier l'affinité d'un ligand pour un G-quadruplexe dans un contexte cellulaire, mais tiendrait également compte de la capacité du ligand à traverser la membrane cellulaire et à pénétrer dans la cellule.

Les systèmes d'expression génique présentés ci-dessus incluant les GQRz pourraient être utilisés comme point de départ dans cette optique. Tout d'abord, il faudrait étudier l'impact de certains ligands déjà bien établis concernant leur spécificité pour la structure G-quadruplexe, sur l'activité des différents GQRz en *cis* *in vitro*. Après confirmation que ces ligands sont capables de moduler l'expression des GQRz, il sera possible de les cloner dans le 5'- ou 3'-UTR du gène rapporteur de la GFP afin de produire les systèmes d'expression génique *in cellulo* (Discussion, Figure 3). Une validation des différents systèmes devrait être effectuée en utilisant les mêmes ligands. Différents contrôles devront être élaborés, *in vitro* et *in cellulo*, afin de s'assurer que l'impact des ligands provient de leur effet sur la structure G-quadruplexe et l'activité du ribozyme. Si les résultats sont encourageants et reproductibles, ces systèmes pourraient être utilisés à grande échelle (p. ex. dans des plaques 96 puits) dans le but de tester l'efficacité d'une pléthore de ligands de G-quadruplexe dans un contexte cellulaire.

En conclusion, les résultats obtenus avec le GQRz auront enrichi nos connaissances sur la capacité d'une structure G-quadruplexe à réguler une activité de l'ARN (Beaudoin and Perreault, 2008). Son mécanisme d'action initial et la possibilité de le modifier efficacement par ingénierie moléculaire démontrent bien toute la flexibilité et le potentiel de cette structure comme élément régulateur de l'ARN. De plus, les outils développés pourraient être profitables pour des études futures portant sur les G-quadruplexes d'ARN et leurs rôles dans la biologie de la cellule. Cette première étude m'a permis d'attaquer avec confiance le second projet de ma thèse, celui porté sur l'étude des G-quadruplexes naturels présents dans les 5'-UTR des ARNm chez l'humain.

3. Étude des G-quadruplexes dans les 5'-UTR

Lors de la publication de l'article présenté au chapitre 2 (Beaudoin and Perreault, 2010), il y avait déjà six publications concernant des G-quadruplexes présents dans des 5'-UTR capables d'agir comme répresseurs traductionnels. La première

d'entre elles était le travail pionnier du laboratoire du Dr Shankar Balasubramanian qui démontra pour la première fois ce phénomène *in vitro* en extrait de réticulocytes (Kumari et al., 2007). Par la suite, deux équipes étudièrent des G-quadruplexes artificiels afin d'en apprendre davantage sur les règles régissant ce phénomène chez la bactérie E. *coli* et, ensuite, chez les cellules de mammifères (Wieland and Hartig, 2007; Halder et al., 2009b). La première preuve de ce phénomène en cellule de mammifère appartient toutefois au laboratoire du Dr Jens Kurreck et portait sur l'étude du G-quadruplex présent dans le 5'-UTR du gène *Zic-1* (Arora et al., 2008). Finalement, deux autres études rapportèrent la présence de deux nouvelles structures G-quadruplexes, agissant comme répresseurs traductionnels dans les cellules de mammifères, dans les 5'-UTR des ARNm des gènes *MT3-MMP* et *ESR1* (Balkwill et al., 2009; Morris and Basu, 2009). Malgré l'existence d'une littérature respectable sur le sujet, notre étude permit d'en apprendre davantage sur les G-quadruplexes présents dans les 5'-UTR et se démarqua des autres grâce à la présence de plusieurs différences innovatrices.

3.1 Une vue globale du phénomène

Un des buts initiaux de ce projet était de développer une méthodologie permettant d'évaluer et de caractériser l'impact général des G-quadruplexes dans les 5'-UTR agissant comme répresseurs traductionnels. C'est pourquoi, dans le but d'avoir une vue plus globale du phénomène, nous avons utilisé une approche séquentielle en trois étapes distinctes: *in silico*, *in vitro* et *in cellulo*. De plus, nous avons étudié neuf candidats en simultanés. L'ajout d'une partie bioinformatique étoffée (*in silico*) était primordial pour avoir une vue d'ensemble comme mentionné au chapitre 2. L'un des résultats bioinformatiques les plus intriguants, selon moi, est la distribution asymétrique des séquences formant potentiellement des G-quadruplexes (G4P pour G-quadruplexe potentiel) entre celles retrouvées sur le brin complémentaire et celles sur le brin matrice de l'ADN (Chapitre 2, Table 1). En effet, pour notre banque de 5'-UTR humains, près de 60% des G4P se sont retrouvés sur le brin matrice alors que seulement 40% sur le brin complémentaire.

Cette observation a également été rapportée précédemment par une autre étude (Huppert et al., 2008). L'explication la plus commune à ce sujet est la suivante: "Puisque les G-quadruplexes présents sur le brin complémentaire se retrouvent dans l'ARNm et peuvent agir comme répresseurs traductionnels, l'appauvrissement de G4P sur le brin complémentaire serait un résultat évolutif. La cellule ayant une tendance évolutive à limiter la présence de G-quadruplexe dans le 5'-UTR de ses ARNm.". Dans le futur, il serait intéressant d'étudier l'impact des G-quadruplexes formés dans le brin matrice d'ADN et situés en aval du site d'initiation de la transcription. Il est fort possible que ce motif d'ADN possède un rôle dans la régulation de la transcription et de l'expression génique. Malheureusement, très peu d'études se sont attardées à ces G-quadruplexes présents dans le brin matrice d'ADN des 5'-UTR. La même méthodologie pourrait être utilisée à cette fin, mais cette fois, en considérant les G4P présents sur le brin matrice (c.-à-d. correspondant à des séries de cytosines dans l'ARNm). En effet, il a été rapporté récemment que l'ADN polymérase ADN dépendante, impliquée dans la réplication du génome, avait besoin d'une hélicase particulière (Pif1), connue pour défaire les structures G-quadruplexes, afin de répliquer fidèlement les régions du génome susceptibles à la formation de cette structure (Lopes et al., 2011; Paeschke et al., 2011). À la lumière de ces résultats, il serait intéressant de vérifier si l'ARN polymérase II ne serait pas affectée par la présence de G-quadruplexes d'ADN sur le brin matrice et si elle ne possèderait pas un mécanisme similaire pour y faire face.

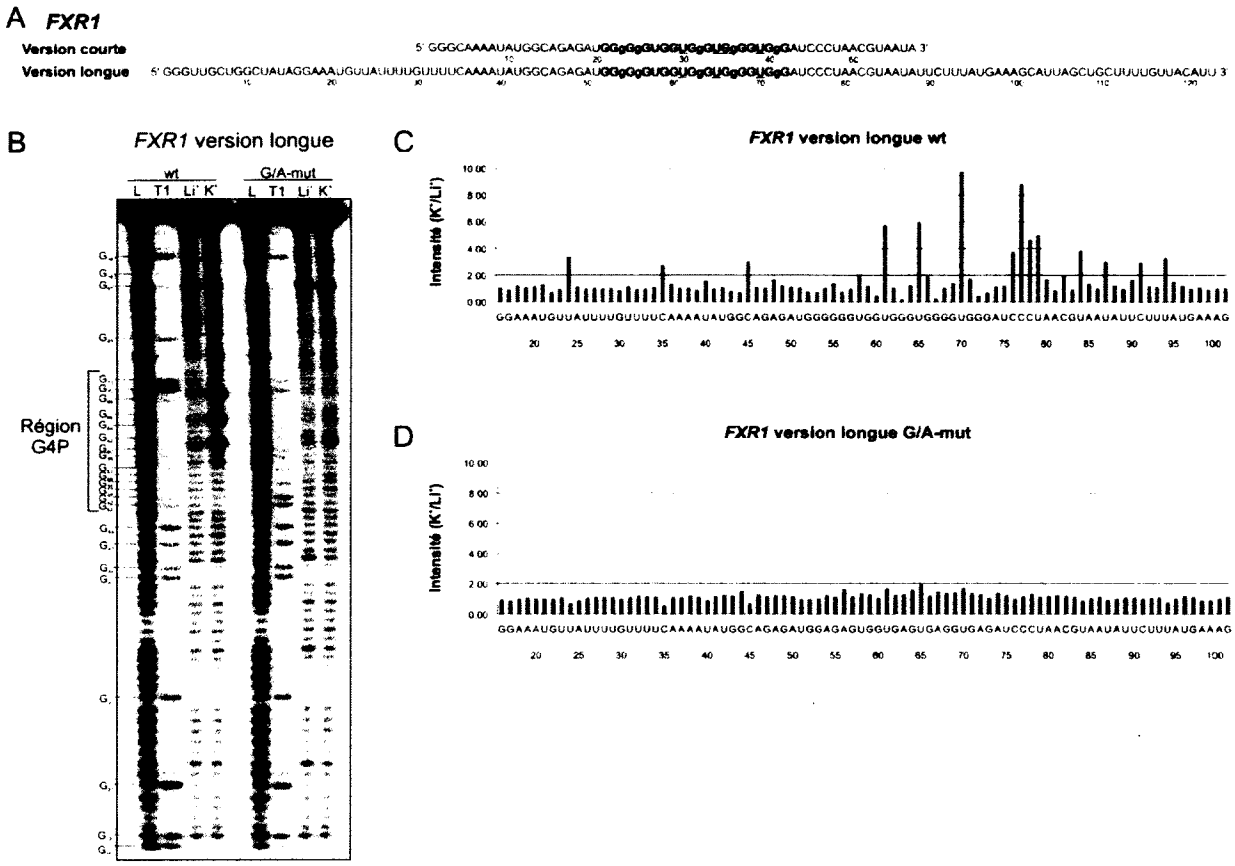
En plus de nous informer sur la fréquence et la distribution des G4P dans les 5'-UTR, l'analyse *in silico* nous a permis de construire un échantillonnage de neuf candidats, représentant des gènes impliqués dans différentes fonctions et processus cellulaires. Les séquences de ces neuf G4P furent testées pour leur capacité à former une structure G-quadruplexe *in vitro*. Six de ces neuf séquences (67%) pouvaient se replier en une structure G-quadruplexe à une concentration physiologique de K^+. La capacité de chacun de ces six G-quadruplexes à réprimer la traduction d'un gène rapporteur a été évaluée *in cellulo* dans le contexte de leur

5'-UTR complet. Ces six candidats (100%) ont tous démontré une répression de la traduction variant de 1.6 à 2.5 fois par rapport à une version mutante du G-quadruplexe (Chapitre 2, Table 4). En tenant compte de ses deux ratios et du nombre de G4P identifiés dans les 5'-UTR des ARNm humains (7198), on obtient un ordre de grandeur de structures G-quadruplexes dans les 5'-UTR de 4798. Évidemment, ce chiffre est loin d'être absolu et de représenter le nombre de G-quadruplexes réels dans ces régions du transcriptome. Toutefois, il met clairement en évidence que les G-quadruplexes dans les 5'-UTR, agissant comme répresseurs traductionnels, sont probablement bien répendus dans la cellule et non seulement limités à quelques exemples éparpillés. C'est vraisemblablement le message le plus important véhiculé par cette étude, puisqu'un échantillon sélectionné pour posséder le moins de biais possible (c.-à-d. constitué de G-quadruplexes impliqués dans différents phénomènes cellulaires, de longueurs différentes, incluant des boucles 1 à 3 de longueurs et séquences différentes, faisant partie de 5'-UTR de différentes longueurs et structures et situés à différentes distances de la structure coiffe de l'ARNm et du codon d'initiation de la traduction AUG) a démontré que 67% des séquences représentaient des G-quadruplexes capables de réprimer la traduction. Finalement, une méthodologie composée de trois approches complémentaires (*in silico*, *in vitro* et *in cellulo*) et l'étude de plusieurs candidats étaient nécessaires afin d'arriver à ces conclusions.

3.2 Le "in-line probing" pour étudier la formation de G-quadruplexes

Cette technique ne demandant aucun produit spécifique ou coûteux, comme des enzymes ou des réactifs, est probablement l'une des méthodes de cartographie les plus simplistes qui existe, comme il a été mentionné dans l'introduction. Bien qu'elle a été largement utilisée pour cartographier les riborégulateurs, cette technique n'avait jamais été mise à profit dans le but d'étudier des structures G-quadruplexes. Pourtant, elle s'avère être une méthode de choix pour différentes raisons.

Tout d'abord, l'analyse par "*in-line probing*" se fait avec l'aide d'ARN radiomarqué au P^{32}, ce qui permet l'utilisation de très faible concentration d'ARN (<1 nM) favorisant les interactions intramoléculaires et limitant énormément la formation de structures intermoléculaires. C'est l'une des différences majeures avec les autres techniques traditionnellement utilisées pour l'étude des G-quadruplexes, qui nécessitent des concentrations dans l'ordre du bas micromolaire (µM). De plus, le "*in-line probing*" permet de suivre aisément les changements de structures observés entre des conditions défavorables à la formation de G-quadruplexes (absence de contre-ion et en présence de Li^+) et celles favorables à cette dernière (présence Na^+ et K^+). Le cas échéant, la formation d'un G-quadruplexe est caractérisée par une augmentation importante du clivage au niveau des nucléotides qui se retrouvent dans les boucles de la structure puisqu'ils deviennent davantage simple brin et gagnent en flexibilité (Chapitre 2, Figure 1-3 et Chapitre 3, Figure 1). Voici un bref résumé des différentes étapes de l'analyse par "*in-line probing*" en utilisant comme exemple le G-quadruplexe retrouvé dans le 3'-UTR de *FXR1* (voir Chapitre 3). Le gel de cartographie de *FXR1* version longue présenté à la Discussion, Figure 6 est une gracieuseté de madame Rachel Jodoin. Tout d'abord, l'efficacité de la technique du "*in-line probing*" à étudier une structure G-quadruplexe présente dans une longue molécule d'ARN a été testée. Plus précisément, cinquante nucléotides en amont et en aval de la séquence correspondant au G4P ont été ajoutés (100 nt au total). Dans le cas de *FXR1*, l'ARN version allongée correspond à une molécule d'une longueur de 124 nucléotides, dont la séquence est indiquée à la Discussion, Figure 6. Une version mutante du G-quadruplexe a également été synthétisée à titre de contrôle. L'étape du "*in-line probing*" a été réalisée de la même façon que mentionnée précédemment et les résultats de la cartographie ont été séparés sur gel dénaturant PAGE (Discussion, Figure 6B). Par la suite, l'intensité de chaque bande du gel a été quantifiée en utilisant le logiciel SAFA ("*Semi-Automated Footprinting Analysis*")(Laederach et al., 2008). Pour chaque bande, un ratio de l'intensité en présence de K^+ divisée par l'intensité en présence de Li^+ a été calculé.



Discussion, Figure 6. Analyse de G-quadruplexe par "*in-line probing*".

(**A**) Séquences primaires des versions courte et longue du candidat *FXR1* utilisées pour l'analyse par "*in-line probing*". La séquence correspondant au G4P est encadrée en jaune. Les guanines en minuscules représentent celles substituées en adénines dans les versions G/A-mut. Les nucléotides soulignés correspondent aux positions devenant significativement plus accessibles en présence de K^+ *versus* en présence de Li^+ après quantification. (**B**) Autoradiogramme typique d'un gel de cartographie par "*in-line probing*" des versions longues de *FXR1* wt et G/A-mut radiomarquées en 5' en présence de 100 mM soit de Li^+ ou de K^+. Les lignes L et T1 représentent respectivement des échelles moléculaires d'hydrolyse alcaline et à la ribonucléase T1. La région du G4P et la localisation de certaines guanines sont représentées à la gauche du gel. (**C,D**) Histogrammes des ratios d'intensités pour chaque nucléotide des versions longues *FXR1* wt (**C**) et *FXR1* G/A-mut (**D**). L'axe des Y représente le ratio des intensités obtenues pour une même bande en présence de K^+ par rapport à en présence de Li^+. L'axe des X montre la séquence et identifie chacune des positions.

Généralement, chaque gel est produit en duplicata et la moyenne des ratios de chaque bande est utilisée pour tracer un graphique d'intensité (ratio K^+/Li^+) en fonction de chaque nucléotide de la séquence étudiée (Discussion, Figure 6C,D). Cette représentation graphique permet d'identifier rapidement les nucléotides qui deviennent plus accessibles en présence de K^+ comparativement à en présence de Li^+. Ils correspondent aux positions possédant une valeur plus élevée à un seuil déterminé (p. ex. 2 dans ce cas-ci) et, dans le cas de *FXR1* wt, ils corrèlent très bien avec les nucléotides retrouvés dans les boucles prédites de la structure G-quadruplexe. Les valeurs recueillies pour *FXR1* G/A-mut sont toutes inférieures à deux, supportant l'incapacité de cette séquence à former un G-quadruplexe. Lorsque de tels types de graphiques sont obtenus, cela suggère fortement que la séquence étudiée est capable de former un G-quadruplexe intramoléculaire en présence d'une concentration physiologique de K^+ *in vitro*. Il est à noter que pour des molécules d'ARN de cette longueur, l'analyse des gels ne permet pas d'avoir de l'information assez précise sur les premiers et derniers 10-20 nucléotides. En résumé, il s'avère très facile de mettre à profit cette technique simple de cartographie de structures d'ARN qu'est le "*in-line probing*" dans le but d'étudier de manière rapide, efficace, informative et reproductible la formation de G-quadruplexe. L'étude présentée au chapitre 2 l'a démontré pour la première fois.

3.3 L'ajout de séquences adjacentes lors de l'étude in vitro

Une autre nouveauté de cette étude (Chapitre 2) a été l'utilisation de séquence d'ARN excédant d'environ 10 à 15 nucléotides en 5' et en 3' la région correspondant au G4P pour l'étude *in vitro*. En effet, la longueur moyenne des G4P pour cette étude était de 23 nucléotides alors que celle des versions d'étude *in vitro* était de 50 nucléotides. Le candidat *FZD2* démontre bien la logique de cet ajout. Sa séquence G4P est composée de 25 nucléotides dont 21 sont des guanines (Chapitre 2, Figure 1A). L'étude *in vitro* limitée au G4P serait pratiquement comme étudier la capacité d'un oligonucléotide poly-guanines à former un G-quadruplexe, expérience réalisée en 1974 (Arnott et al., 1974). La

nature des nucléotides ajoutés en 5' et en 3' est celle des séquences retrouvées dans le contexte du 5'-UTR naturel de chacun des différents candidats. Les résultats enregistrés *in vitro* sont donc plus représentatifs de la réalité et des contraintes du contexte génique présent *in cellulo*. Cette modification au protocole standard nous a permis de mettre le doigt sur une nouvelle règle régissant l'adoption d'une structure G-quadruplexé par les molécules d'ARN.

Comme il a été mentionné précédemment, des neuf candidats étudiés initialement, trois n'étaient pas en mesure d'adopter une structure G-quadruplexe en présence de condition physiologique de K^+ *in vitro* (Chapitre 2, Table 4). Suite à une analyse de la séquence primaire des versions *in vitro*, la présence de plusieurs séries de cytosines a été identifiée dans les séquences adjacentes au G4P pour ces trois candidats comparativement aux six autres. La présence de ces cytosines générait des ratios G/C plus bas, s'approchant de 1, pour les trois candidats incapables de former un G-quadruplexe *in vitro* (Chapitre 2, Table 5). La composition riche en guanines et cytosines, de ces candidats, augmente de beaucoup la capacité de ces molécules d'ARN à adopter des structures secondaires très stables impliquant plusieurs paires de bases Watson-Crick G-C. Ce fait est supporté par l'obtention de valeurs d'énergie plus basses des structures secondaires prédites pour ces trois candidats négatifs en utilisant trois programmes différents de prédictions de structures (Chapitre 2, Table 5). La formation de structures secondaires inhibitrices plus stables, échafaudée via des paires de bases Watson-Crik, semble empêcher la liberté des guanines à interagir entre-elles et à former une structure G-quadruplexe. Supportant cette hypothèse, l'analyse de mutants, où ces structures secondaires inhibitrices ont été affaiblies par la substitution de plusieurs cytosines pour des adénosines, a démontré que ces modifications avaient le pouvoir de transformer chacun de ces candidats négatifs en positifs. Cette nouvelle capacité à former une structure G-quadruplexe à la suite de ces mutations a été confirmée non seulement *in vitro* pour ces trois candidats, mais également *in cellulo* pour *TNFSF12* et son aptitude à réprimer la traduction.

La découverte, que la présence de séries de cytosines dans les régions adjacentes au G-quadruplexe peut interférer et réguler la formation de cette structure, ajoute littéralement un nouveau niveau de régulation possible. Certaines protéines inattendues peuvent maintenant être considérées comme ayant un impact potentiel sur la régulation et la formation de G-quadruplexes d'ARN *in cellulo*. Par exemple, des protéines du groupe des "*poly(rC)-binding protein*" (PCBP), possédant une affinité particulière pour les séries de cytosines, pourraient déstabiliser une structure secondaire inhibitrice et permettre la formation de G-quadruplexes à la suite de leurs liaisons. Certaines hélicases, capables de défaire des structures secondaires formées de paires de bases Watson-Crick, auraient également le potentiel de générer le même effet grâce à leur activité catalytique (Chapitre 2, Figure 4). En résumé, le contexte génique dans lequel la structure G-quadruplexe se retrouve, c'est-à-dire la stabilité globale des structures secondaires Watson-Crick adoptées dans cette région, semble être un facteur primordial afin de déterminer si un G-quadruplexe sera formé ou non dans un transcrit spécifique. Par la suite, deux questions importantes peuvent être soulevées. Combien de nucléotides en 5' et en 3' du G-quadruplexe devraient être considérés à cet effet et quels paramètres (p. ex. le ratio G/C ou l'énergie prédite des structures secondaires) devraient être examinés afin de prédire la formation ou non d'un G-quadruplexe dans une région donnée? Ces deux questions seront approfondies plus loin dans cette discussion.

3.4 La présence de SNP dans les G-quadruplexes

La banque de données de SNP au niveau de G4P rapportée dans cette étude et appuyée par des résultats *in silico*, *in vitro* et *in cellulo* est également un fait d'arme de ce travail. La démonstration qu'un SNP présent dans le G-quadruplexe *AASDHPPT* soit capable de fortement déstabiliser cette structure d'ARN et d'empêcher sa formation *in vitro* ainsi que de diminuer sa capacité à réprimer la traduction *in cellulo*, constitue une preuve de concept très intéressante sur l'importance et l'impact que peuvent avoir ces SNP sur ces structures et

l'expression génique. Précédent cette étude, une recherche bioinformatique a analysé le polymorphisme des motifs G-quadruplexes à la grandeur du génome humain. Elle suggérait que les positions menant à l'abolition d'un G-quadruplexe (les séries de guanines) étaient moins polymorphes et plus conservées que leurs homologues neutres (sans impact sur la structure, p. ex. dans les boucles)(Nakken et al., 2009). Tout récemment, une autre étude similaire à la nôtre, comprenant une partie *in silico*, *in vitro* et *in cellulo*, a démontré que la présence de SNP pouvait influencer l'expression génique par la déstabilisation des G-quadruplexes d'ADN retrouvés dans les promoteurs (Baral et al., 2012). Ces données sont très intéressantes et importantes car il semble de plus en plus évident que plusieurs SNP reliés et associés à des maladies peuvent affecter l'homéostasie de la cellule en altérant des structures d'ARN (Halvorsen et al., 2010). Finalement, avec le nombre de rôles grandissant concernant la structure G-quadruplexe et son importance dans l'expression génique, il ne serait pas surprenant d'éventuellement trouver des désordres biologiques causés par la présence de SNP déstabilisant cette structure. La banque de données et les résultats présentés ici peuvent constituer un point de départ pour de telles recherches.

4.0 Étude des G-quadruplexes dans les 3'-UTR

La même méthodologie en trois étapes (*in silico*, *in vitro* et *in cellulo*) développée ultérieurement pour le projet des 5'-UTR a été utilisée pour l'étude des G-quadruplexes présents dans les 3'-UTR. À l'inverse des 5'-UTR, où un rôle relativement bien établi existait pour la structure G-quadruplexe, les évidences de rôles pour cette structure dans les 3'-UTR se font toujours attendre. Dans l'étude présentée au chapitre 3, l'approche utilisée a permis d'amasser une quantité importante de renseignement sur la présence et la dispersion des G4P présents dans les 3'-UTR humains. Probablement plus important encore, elle a permis d'identifier deux G-quadruplexes présents dans deux 3'-UTR différents et de caractériser leur impact et leur importance sur l'expression génique. Si l'étude de ces deux candidats suggérait un rôle pour la structure G-quadruplexe à stimuler

l'utilisation d'un site de polyadénylation alternatif, situé à l'intérieur du 3'-UTR, une analyse plus poussée de celle retrouvée dans l'ARNm du gène *FXR1* a laissé entrevoir deux autres rôles possibles très intéressants et intrigants.

4.1 Les G-quadruplexes et la polyadénylation alternative

Tout d'abord, la polyadénylation alternative des ARNm a récemment été démontrée comme étant un mécanisme général dans la cellule et non pas limitée à quelques exceptions (Di Giammartino et al., 2011). Toutefois, il est toujours difficile de déterminer quels sites de polyadénylation seront utilisés dans les ARNm à un moment précis, ainsi que de savoir comment les conditions cellulaires peuvent affecter ces choix. Afin d'arriver à ce niveau de connaissance, il faut identifier et caractériser davantage les éléments impliqués dans la régulation de ces différents sites de polyadénylation. Comme il a été rapporté dans l'étude présentée au chapitre 3 de cette thèse, les structures G-quadruplexes présentes dans les 3'-UTR semblent pouvoir agir comme éléments post-transcriptionnels en *cis* stimulant l'utilisation d'un site de polyadénylation situé en amont (Chapitre 3). Brièvement, ce phénomène a été démontré pour deux G-quadruplexes différents situés dans deux 3'-UTR distincts. Dans le cas du G-quadruplexe *LRP5*, le site de polyadénylation alternatif (situé à l'intérieur du 3'-UTR) est le seul site actif à cet effet, puisque le site canonique (identifié par le "*National Center for Biotechnology Information*", NCBI) est inactif dans les constructions utilisées. Alors, la modulation du niveau de polyadénylation du site alternatif, par la présence ou l'absence du G-quadruplexe, dicte la quantité d'ARNm produit et, par le fait même, la quantité de protéines synthétisées (Chapitre 3, Figure 2). Dans le cas du G-quadruplexe *FXR1*, son 3'-UTR possède un site alternatif et un site canonique actifs menant à la production d'un isoforme court, via l'utilisation du site alternatif, et d'un isoforme long, via l'utilisation du site canonique (Chapitre 3, Figure 3). La quantité totale d'ARNm produit en présence ou en absence du G-quadruplexe ne change donc pas significativement. Une différence est plutôt observée au niveau du ratio de l'accumulation de l'isoforme court *versus* l'isoforme long. La présence du G-

quadruplexe favorisant l'accumulation de l'isoforme court par rapport au long alors que cette tendance se renverse en absence de cette structure (accumulation de l'isoforme long comparativement au court)(Chapitre 3, Figure 3). La capacité du G-quadruplexe à stimuler la polyadénylation au site aternatif et à favoriser l'accumulation de l'isoforme court, possédant un 3'-UTR raccourci contenant moins de sites de liaisons pour les miARN, mène à une augmentation du niveau de traduction et de synthèse protéique du gène rapporteur (Chapitre 3, Figure 4). En résumé, l'impact au niveau de la traduction, causé par la structure G-quadruplexe *FXR1*, provient majoritairement de la perturbation du réseau de régulation par les miARN, alors que pour *LRP5*, il est presque entièrement provoqué par les différents niveaux d'ARNm total produit. Les mécanismes par lesquels une structure G-quadruplexe pourrait stimuler la polyadénylation alternative d'un site en amont sont discutés dans le chapitre 3. Cette étude met en évidence deux phénomènes par lequel la stimulation de l'efficacité d'un site de polyadénylation alternatif par une structure G-quadruplexe peut mener à un changement au niveau de l'expression génique.

4.2 Les G-quadruplexes et l'inhibition d'un site de polyadénylation en aval

Par la suite, la caractérisation de la structure G-quadruplexe présente dans le 3'-UTR de l'ARNm *FXR1* a permis de soulever plusieurs autres faits intéressants. Outre la stimulation de la polyadénylation au niveau du site alternatif, situé en amont du G-quadruplexe, la présence du G-quadruplexe semblait diminuer la quantité d'ARNm polyadénylé au site canonique, situé en aval de cette structure (Chapitre 3, Figure 3B). A priori, une explication logique à ce phénomène était: si la présence du G-quadruplexe menait à une efficacité de polyadénylation accrue au niveau du site alternatif, pour une même quantité d'ARNm transcrit, une perte d'activité à ce niveau, par la mutation du G-quadruplexe, entrainerait une augmentation de la polyadénylation au site canonique. C'est ce qui a été observé (Chapitre 3, Figure 3B). Toutefois, lorsque le signal de polyadénylation du site alternatif est aboli (AltPAS-mut, voir Chapitre 3), une diminution de la quantité

totale d'ARNm polyadénylés en présence du G-quadruplexe était toujours observée. À l'inverse, l'absence de cette structure permettait un gain à ce niveau (Chapitre 3, Figure 3C). Le fait que le phénotype constaté au niveau du site canonique était indépendant de l'efficacité du site alternatif venait réfuter cette première hypothèse. Il est intéressant de noter que la différence d'ARNm totaux produits en présence ou en absence du G-quadruplexe corrèle très bien avec les niveaux d'ARNm polyadénylés au niveau du site canonique pour les constructions AltPAS-mut (comparer Chapitre 3, Figure 3E et 3D). Cette observation suggère que l'isoforme long est le seul transcrit stable produit par ces constructions. Il a été rapporté dans la littérature que certaines régions riches en guanines ont la capacité de former une structure particulière appelée "*R-loop*" (transcrit nouvellement synthétisé formant un hétéroduplexe ARN-ADN avec le brin d'ADN matrice par appariement Watson-Crick) et seraient présentes dans environ 59% des transcrits (Wongsurawat et al., 2012). Lorsque présent dans la région 3' d'un gène, une "*R-loop*" peut agir comme un site de pause pour l'ARN polymérase II et favoriser la terminaison de la transcription (Skourti-Stathaki et al., 2011). Une autre étude récente a démontré qu'il y avait la formation d'une "*R-loop*" entre le transcrit riche en guanines et le brin complémentaire également riche en guanines, lors de la transcription de la région CSB II ("*Conserved Sequence Block II*") de l'ADN mitochondrial. Cette "*R-loop*" serait en fait une structure G-quadruplexe bimoléculaire formée entre le brin d'ARN et d'ADN. La formation de ce G-quadruplexe hybride a été suggérée pour réguler la formation des amorces d'ARN nécessaire afin d'initier la réplication du génome mitochondrial (Wanrooij et al., 2012). Finalement, une autre étude a rapporté que la formation d'une structure G-quadruplexe, dans un transcrit nouvellement synthétisé dans la mitochondrie, pouvait stimuler la terminaison de la transcription (Wanrooij et al., 2010). En somme, il semble de plus en plus évident qu'une région riche en guanines est capable de promouvoir la terminaison de la transcription et différents mécanismes ont été suggérés. Une augmentation de la terminaison de la transcription au niveau de la séquence formant un G-quadruplexe, dans le 3'-UTR de *FXR1*, pourrait expliquer la diminution d'ARNm produit au site de polyadénylation

canonique. Les résultats obtenus jusqu'à présent ne permettent pas de discriminer quels mécanismes semblent être à l'œuvre dans ce contexte. Il serait intéressant de vérifier l'impact du ligand spécifique à la structure G-quadruplexe, PhenDC3, sur le phénotype observé au niveau du site canonique. Les résultats recueillis pourraient nous éclairer à savoir si ce phénomène est en partie causé par la formation d'une structure G-quadruplexe. Cependant, ils ne permettraient pas de déterminer de quel type de G-quadruplexe il s'agit: l'hybride bimoléculaire, formé entre le transcrit et le brin d'ADN complémentaire, ou l'intramoléculaire d'ARN, formé dans l'ARNm seulement. Pour conclure, certains des résultats obtenus suggèrent un second rôle pour la séquence formant le G-quadruplexe *FXR1* et il serait très intéressant d'approfondir cette nouvelle piste dans le futur.

4.3 Les G-quadruplexes comme activateurs de la traduction?

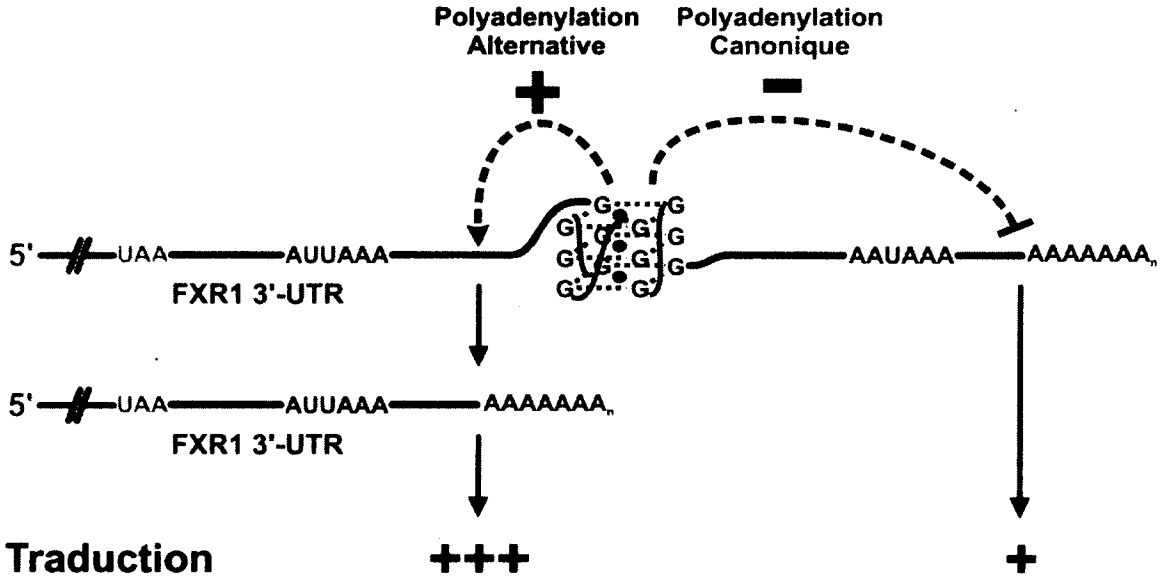
Finalement, une observation additionnelle et curieuse a été notée lors de la caractérisation du G-quadruplexe présent dans le 3'-UTR de l'ARNm du gène *FXR1*. Une fois de plus, c'est l'étude des constructions AltPAS-mut qui offre les résultats préliminaires suggérant ce nouveau rôle pour une structure G-quadruplexe présente dans un 3'-UTR. Comme il a été discuté dans la section précédente, la présence du G-quadruplexe *FXR1* mène à une diminution de la quantité d'ARNm produits au site de polyadénylation canonique et, ce, indépendamment de l'efficacité de la polyadénylation au niveau du site alternatif. En ce qui concerne les constructions AltPAS-mut, uniquement l'isoforme long est synthétisé et dicte du même coup la quantité globale d'ARNm produits. Les degrés d'expression génique au niveau de l'ARN pour les constructions AltPAS-mut ont démontré qu'il y avait plus d'ARNm produits en absence du G-quadruplexe qu'en présence de la structure. Cette conclusion a été corroborée par deux techniques différentes de quantification d'ARN, c'est-à-dire par buvardage northern couplé à la ribonucléase H et par essai de protection aux ribonucléases (Chapitre 3, Figure 3D,E). Toutefois, l'expression génique au niveau protéique, par essai luciférase, a indiqué qu'une plus grande quantité de protéines était synthétisée en présence du

G-quadruplexe qu'en absence de cette structure (Chapitre 3, Figure 3F). Conséquemment, pour les constructions AltPAS-mut, une plus petite quantité d'ARNm contenant un G-quadruplexe a mené à la synthèse d'un plus grand nombre de protéines qu'une plus grande quantité d'ARNm ne possédant pas de G-quadruplexe. Ces résultats suggèrent que la présence d'une structure G-quadruplexe dans le 3'-UTR d'un ARNm procurerait à celui-ci un meilleur niveau de synthèse protéique. À nouveau, les résultats accumulés concernant ce nouveau rôle potentiel sont préliminaires et une nouvelle série d'expériences sera nécessaire afin de l'étudier plus en détail. Une expérience relativement simple et rapide serait d'utiliser les constructions *FXR1* AltPAS-mut pour synthétiser les ARNm (coiffés et polyadénylés) correspondant aux isoformes longs *in vitro*. Les versions wt et G/A-mut du G-quadruplexe de ces ARNm pourraient être transfectées dans les cellules directement sous forme d'ARN et les niveaux de protéines synthétisées enregistrés quelques heures après la transfection. Cette approche permettrait de comparer les niveaux de synthèse protéique de ces différentes constructions de façon indépendante des phénomènes de maturation cellulaires réalisés au niveau de la transcription et du pré-ARNm. La même méthodologie pourrait être effectuée en absence et en présence du ligand de G-quadruplexes PhenDC3, afin de mettre en évidence l'importance de la structure G-quadruplexe dans ce processus. Un G-quadruplexe présent dans un 3'-UTR pourrait mener à une augmentation de la synthèse protéique de différentes façons. Tout d'abord, il pourrait représenter un site de liaison pour une protéine favorisant l'association de l'ARNm aux polyribosomes. En effet, il a été rapporté que la protéine FMRP est capable de lier, *in vitro*, des séquences G-quadruplexes retrouvées dans les 3'-UTR des ARNm des gènes *Sem3F*, *Arginin vasopressin receptor V1a* et *Munc13*, et que leur association aux polyribosomes est diminuée dans les cellules de patients atteints du syndrôme du X fragile exprimant de très faible quantité de FMRP (Darnell et al., 2001). Par la suite, il pourrait stimuler un meilleur niveau de traduction en affectant l'accessibilité de certains sites de liaison de miARN (Kedde et al., 2010). De plus, certaines structures G-quadruplexes ont déjà été rapportées pour faire partie d'éléments importants de motifs IRES dans

les 5'-UTR permettant la traduction de messagers indépendamment de la coiffe (Bonnal et al., 2003). Alors, il serait possible que cette structure puisse recruter certains facteurs stimulant la traduction même lorsqu'elle est présente dans un 3'-UTR. Avec la circularisation de l'ARNm, le 5'-UTR et 3'-UTR se retrouvent relativement près l'un de l'autre au niveau spatial dans la cellule. Finalement, il serait possible que la structure G-quadruplexe agisse comme élément de localisation afin d'aiguiller l'ARNm dans une région de la cellule plus propice à la traduction. Il a déjà été rapporté que des G-quadruplexes dans les 3'-UTR des gènes *PSD-95* et *CaMKIIa* peuvent agir comme signal de localisation à une région spécifique de la cellule neuronale, aux neurites (Subramanian et al., 2011). En conclusion, l'étude approfondie du candidat *FXR1* a permis d'obtenir des résultats préliminaires sur un troisième rôle attribuable à la région de la structure G-quadruplexe et, bien que le mécanisme reste encore inconnu, il serait très intéressant de s'y attarder prochainement.

En conclusion, l'impact de la présence de G-quadruplexes dans les 3'-UTR semble plus complexe et dépendre davantage du contexte génomique dans lequel il se retrouve que celui des G-quadruplexes présents dans les 5'-UTR. D'un côté, la structure G-quadruplexe peut stimuler l'efficacité d'un site de polyadénylation situé en amont et, de l'autre, diminuer la quantité d'ARNm produits à un site de polyadénylation situé en aval (Discussion, Figure 7). Comme il a été vu, la présence de la structure G-quadruplexe mène généralement à une augmentation de la traduction qui peut être le résultat de différents phénomènes. Elle peut être causée par une augmentation de la quantité d'ARNm produits, comme c'est le cas pour *LRP5*. Elle peut être occasionnée par une plus grande représentation d'un isoforme possédant un 3'-UTR plus court, incluant moins de sites de liaison de microARN, pour une même quantité globale d'ARNm, comme c'est le cas pour *FXR1*. De plus, la présence d'un G-quadruplexe dans le 3'-UTR de la version mature d'un ARNm semble pouvoir augmenter son niveau de synthèse protéique. Avec tous ces différents effets, il serait loin d'être surprenant que la structure G-quadruplexe recrute différents facteurs protéiques afin de moduler ces

phénomènes. D'ailleurs, une étude a démontré qu'une structure G-quadruplexe située en 3' du gène *p53* était capable de maintenir un niveau adéquat de polyadénylation de cet ARNm à la suite d'un stress cellulaire via l'interaction avec les protéines hnRNP H/F, ce qui supporte cette hypothèse (Decorsière et al., 2011). Il est possible que la structure G-quadruplexe interagisse avec certaines protéines au niveau du noyau afin de produire une partie de ces différents effets, comme par exemple au niveau du choix du site de polyadénylation. De plus, une fois exportée au cytoplasme, elle pourrait lier d'autres facteurs afin de favoriser une meilleure synthèse protéique. Comme mentionné plus haut, nos connaissances sur l'importance des G-quadruplexes dans les 3'-UTR sont encore très pauvres. Toutefois, quelques études réalisées sur le sujet, incluant celle présenté au chapitre 3 de cette thèse, font pointer à l'horizon plusieurs implications importantes de ces structures sur le niveau d'expression génique d'un nombre significatif de gènes.



Discussion, Figure 7. Schéma résumant l'impact sur la polyadénylation du G-quadruplexe présent dans le 3'-UTR de *FXR1*.

5.0 La protéine FXR1 et les G-quadruplexes

La protéine FMRP est probablement une des protéines les plus connues à posséder une affinité particulière pour les structures G-quadruplexes d'ARN. Cette protéine est impliquée dans plusieurs étapes du métabolisme des ARNm. Elle joue un rôle dans le trafic nucléocytoplasmique, le contrôle de la traduction et le transport le long des dendrites dans les neurones de plusieurs ARNm (Bardoni et al., 2006). Récemment, la structure en solution du complexe entre le peptide riche en arginines et glycines (RGG) de la protéine FMRP humaine et l'ARN riche en guanines *sc1* a été résolue par RMN (Phan et al., 2011). Elle suggère fortement que FMRP lierait l'ARN à l'intersection entre une région formant un G-quadruplexe et une autre formant un duplex. Cette observation est en accord avec des résultats *in vitro* et *in cellulo* rapportés lors de l'étude de FMRP et des ARNm avec lesquels cette protéine interagit (Brown et al., 2001; Darnell et al., 2001). Plusieurs des sites de liaisons identifiés pour la protéine FMRP se situent au niveau des 3'-UTR. Bon nombre de ces cibles voient leur association avec les polyribosomes, reflétant leur niveau de traduction, augmenter ou diminuer en absence de protéine FMRP dans la cellule. En résumé, il existe un lien solide entre l'affinité de FMRP pour la structure G-quadruplexe, la présence de G-quadruplexe dans un ARNm et la capacité de FMPR à réguler l'expression génique de plusieurs de ces ARNm.

La protéine FMRP possède également de nombreux partenaires protéique avec lesquels elle peut interagir (Bardoni et al., 2006). La liaison à certains de ces partenaires entraine une perte d'affinité pour la structure G-quadruplexe. C'est le cas pour la protéine FXR1P ("*Fragile X Related Protein 1*") encodée par le gène *FXR1*, celui-là même utilisé pour l'étude du chapitre 3. Il existe plusieurs isoformes de la protéine FXR1P produits par épissage alternatif, mais ils conservent généralement tous le même 3'-UTR comprenant la structure G-quadruplexe préalablement caractérisée (Kirkpatrick et al., 1999). La protéine FXR1P est capable de former un hétérodimère avec la protéine FMRP (Bechara et al., 2007; Melko and Bardoni, 2010). Cet hétérodimère perd alors son affinité pour la structure G-quadruplexe. FXR1P est donc capable de moduler l'affinité de FMRP

pour les G-quadruplexes et de potentiellement modifier le rôle que FMRP joue sur l'expression génique. Sachant cela, il est tentant de proposer un rôle pour FMRP et FXR1P dans la régulation de l'expression du gène *FXR1* via les phénomènes affectés par la structure G-quadruplexe identifiés au chapitre 3. Pour se faire, il serait intéressant de vérifier si la surexpression ou le "*knockdown*" de FMRP affecte l'expression protéique de nos différentes constructions. Outre le niveau de protéine, il serait important de regarder si les ratios (isoformes court *versus* long) varient également dans ces conditions. Il faudrait garder en tête que la liaison de FMRP pourrait affecter un ou plusieurs des rôles reliés à la structure G-quadruplexe. Les résultats obtenus lors de ces expériences préliminaires mettraient la table à une nouvelle série d'expériences permettant de caractériser un mécanisme possible de rétroaction de l'expression génique de *FXR1*, impliquant les protéines FMRP, FXR1P et la structure G-quadruplexe retrouvée dans le 3'-UTR de ce gène.

6.0 Les G-quadruplexes à longue boucle 2

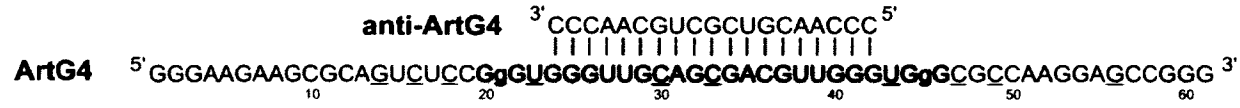
Il a été suggéré qu'une structure G-quadruplexe d'ADN possédant des boucles 1 et 3 très courtes (p. ex. de 1 nucléotide) était capable de supporter la présence d'une très longue boucle 2 (plus longue que 7 nt.) (Guédin et al., 2010). Par exemple, la séquence suivante $G_3TG_3T_{21}G_3TG_3$ est capable de former une structure G-quadruplexe en présence de K^+. À la suite de ces résultats, j'ai voulu m'intéresser aux versions ARN des G-quadruplexes possédant une longue boucle 2. Ce nouveau projet a été séparé en trois parties (résultats non publiés). La première partie était de tester si ces G-quadruplexes d'ARN étaient capables de se former *in vitro* et *in cellulo* en utilisant une séquence artificielle prédéterminée. La deuxième partie était de vérifier s'il était possible d'utiliser cette séquence supplémentaire (c.-à-d. la longue boucle 2) afin de cibler spécifiquement ce G-quadruplexe et de moduler sa formation *in vitro* et *in cellulo*. La troisième partie était de construire une banque de données des G-quadruplexes à longue boucle 2 retrouvés dans les 5'-UTR des ARNm humains et de vérifier s'il était possible de les cibler de cette

façon. En effet, comme il a été mentionné dans l'introduction, il existe un engouement majeur dans la recherche de ligands capables de stabiliser ou de déstabiliser les structures G-quadruplexes. Toutefois, le défi auquel les scientifiques se butent jusqu'à maintenant est celui de pouvoir cibler spécifiquement une topologie ou un G-quadruplexe donné. Ne possédant pas de grande aptitude en chimie, mais voulant participer à ce grand défi, j'ai décidé de me tourner vers une approche davantage axée sur la biologie moléculaire et l'utilisation d'oligonucléotides afin d'arriver à cette fin. Les G-quadruplexes à longue boucle 2 pourraient constituer des cibles de choix afin de mettre au point une telle approche. Pour ce nouveau projet, j'ai travaillé en collaboration avec M. Samuel Rouleau, qui a réalisé la plupart des résultats préliminaires mentionnés dans cette section.

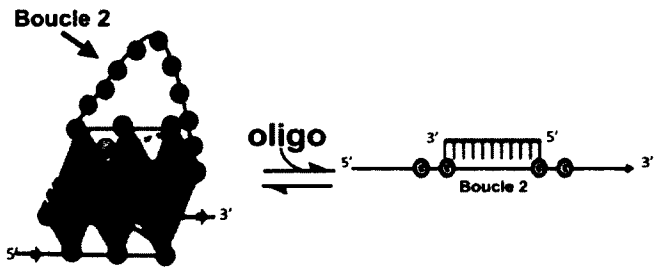
6.1 G-quadruplexe artificiel à longue boucle 2

Tout d'abord, il fallait déterminer si un G-quadruplexe possédant une longue boucle 2 était capable de se former *in vitro*. Un G-quadruplexe artificiel (ArtG4) ayant des boucles 1 et 3 d'un nucléotide et une boucle 2 de 13 nucléotides a été fabriqué (Discussion, Figure 8A). La technique de "*in-line probing*" a été utilisée afin de vérifier la formation d'une structure G-quadruplexe par cette séquence. Les résultats préliminaires obtenues suggèrent fortement que cette séquence, ArtG4, est capable d'adopter une structure G-quadruplexe *in vitro* en présence de K^+. Les versions wt et G/A-mut du G-quadruplexe ArtG4 ont ensuite été clonées dans un système de gènes rapporteurs (le même que celui présenté au Chapitre 2) afin d'étudier leur impact sur l'expression génique. Les résultats préliminaires proposent que le G-quadruplexe ArtG4 est capable de se former *in cellulo* et de réprimer la traduction. En effet, un gain d'environ 2 fois est observé au niveau de la synthèse protéique lorsque la structure G-quadruplexe est abolie. En résumé, les résultats initiaux obtenus *in vitro* et *in cellulo* suggère fortement que cette séquence artificielle est capable d'adopter une structure G-quadruplexe *in vitro* et *in cellulo* ainsi que de réprimer la traduction d'un gène rapporteur *in cellulo*.

A



B



Discussion, Figure 8. Étude du G-quadruplexe artificiel à longue boucle 2.

(**A**) Séquences primaires du *ArtG4* et de l'oligonucléotide anti-ArtG4 utilisées pour l'analyse préliminaire *in vitro*. La séquence encadrée en jaune correspond à la structure G-quadruplexe prédite. Les guanines en minuscule représentent celles substituées en adénine dans la version G/A-mut. Les nucléotides soulignés correspondent aux positions devenant significativement plus accessibles en présence de K^+ *versus* en présence de Li^+ par "*in-line probing*" après quantification. La région d'appariement entre anti-ArtG4 et *ArtG4* est montrée. (**B**) Schéma résumant le mécanisme d'action de l'oligonucléotide anti-G4 permettant d'empêcher ou de défaire la structure G-quadruplexe adoptée par *ArtG4* via sont appariement à la boucle 2 ainsi qu'à la deuxième et troisième séries de guanines.

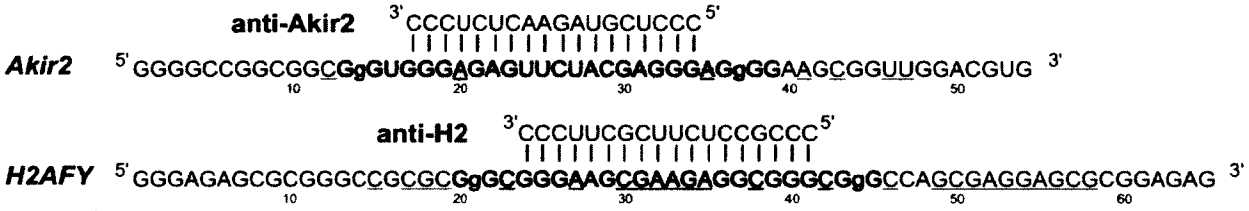
Par la suite, il fallait imaginer une stratégie afin de cibler spécifiquement ce G-quadruplexe avec l'utilisation d'un oligonucléotide. En partant de l'idée d'utiliser la séquence de la longue boucle 2 comme région d'appariement, un oligonucléotide effecteur possédant une séquence complémentaire à cette région a été synthétisé. La complémentarité a également été élargie afin d'inclure la deuxième et la troisième série de guanines du G-quadruplexe (Discussion, Figure 8A,B). Cet oligonucléotide effecteur négatif (anti-ArtG4) devrait donc être capable de reconnaître spécifiquement ce G-quadruplexe, via son appariement au niveau de la boucle 2, et d'empêcher ou de défaire la structure G-quadruplexe par l'interaction avec les séries de guanines 2 et 3 (Discussion, Figure 8A,B). Des expériences préliminaires de "*in-line probing*" ont été réalisées afin d'étudier l'impact du anti-ArtG4 sur la formation du G-quadruplexe. Les résultats recueillis

insinuent que la séquence ArtG4 perd sa capacité à adopter une structure G-quadruplexe en présence du anti-ArtG4. L'impact de la présence du anti-ArtG4 a également été testé de façon préliminaire *in cellulo*. Les résultats préliminaires semblent démontrer que la présence du ArtG4 entraine une diminution de 50% de la répression causée par la structure G-quadruplexe au niveau de la synthèse protéique. Ces résultats, bien que préliminaires, semblent être très prometteurs dans le but de pouvoir cibler spécifiquement un G-quadruplexe donné *in cellulo* et, par le fait même, modifier l'expression génique.

6.2 G-quadruplexes naturels à longue boucle 2

Cette preuve de concept effectuée avec l'aide d'un G-quadruplexe artificiel, il était intéressant d'identifier des cibles naturelles potentielles. Dans cette optique, une analyse bioinformatique des 5'- et 3'-UTR humains a été effectuée. L'algorithme utilisé pour la recherche de G4P dans ces régions des ARNm était le suivant: G_x-H-G_x-$N_{2\text{-}90}$-G_x-H-G_x où $x \geq 3$, N correspond à n'importe lequel des quatre nucléotides et H signifie n'importe lequel des quatre nucléotides à l'exception d'une guanosine. Un total de 1453 G4P et 2282 G4P ont été identifiés dans les banques de 5'- et 3'-UTR respectivement (Discussion, Dataset S1 et Dataset S2, voir Annexe 1). Parmi ceux-ci, 1231 G4P et 1853 G4P possèdent une boucle 2 plus longue que sept nucléotides pour les banques de 5'- et 3'-UTR respectivement. La banque de G4P à longue boucle 2 retrouvés dans les 3'-UTR pourra être utilisée pour des études futures puisqu'il semble très probable que certaines de ses séquences puissent former des G-quadruplexes *in vitro* et *in cellulo*. Cependant, à la suite des résultats préliminaires obtenus précédemment, une attention particulière a été apportée à la banque de G4P à longue boucle 2 retrouvés dans les 5'-UTR. Deux candidats ont été sélectionnés afin de tester leur habilité à former une structure G-quadruplexe *in vitro* et à réprimer la traduction *in cellulo*: *akir2* (*akirin 2*) et *H2AFY* (*H2A histone family member Y*)(Discussion, Figure 9).Les résultats préliminaires obtenus, concernant ces deux candidats possédant une boucle 2 de 12 nucléotides, proposent qu'ils sont capables d'adopter des

structures G-quadruplexes *in vitro* et d'inhiber l'expression génique *in cellulo*. De plus, la présence de leur effecteur spécifique respectif (anti-Akir2 et anti-H2) semble inhiber en partie la formation de G-quadruplexe par ces séquences *in vitro* et *in cellulo* (Discussion, Figure 9). En conclusion, l'approche développée à base d'oligonucléotides pour cibler spécifiquement un G-quadruplexe donné semble être prometteuse et mérite d'être étudiée davantage dans le futur. Une approche, basée sur la reconnaissance de séquences adjacentes ou internes à un G-quadruplexe, pourrait bien être une alternative de choix aux petites molécules chimiques dans le but de cibler un G-quadruplexe précis parmi la pléthore retrouvée *in cellulo*. Pour y arriver par contre, nos connaissances sur les paramètres régissant leur formation *in cellulo* doivent continuer de s'étoffer.



Discussion, Figure 9. Étude des G-quadruplexes naturels à longue boucle 2 *akir2* et *H2AFY*.

Séquences primaires des G-quadruplexes *akir2* et *H2AFY* et des oligonucléotides anti-Akir2 et anti-H2 utilisées pour l'analyse préliminaire *in vitro*. Les séquences encadrées en jaune correspondent aux structures G-quadruplexes prédites. Les guanines en minuscule représentent celles substituées en adénine dans la version G/A-mut. Les nucléotides soulignés correspondent aux positions devenant significativement plus accessibles en présence de K^+ *versus* en présence de Li^+ par "*in-line probing*" après quantification. Les régions d'appariement entre anti-Akir2 et *akir2* et entre anti-H2 et *H2AFY* sont montrées.

7. Le contexte génique et la formation de G-quadruplexe

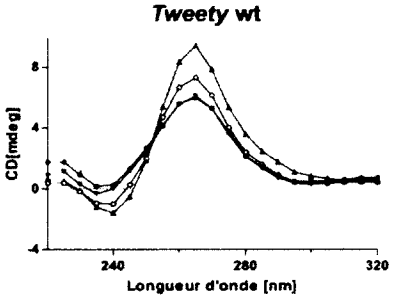
Certains des résultats présentés au chapitre 2 démontrent que les séquences adjacentes au G-quadruplexe et le contexte génique peuvent être cruciaux pour la formation de cette structure *in vitro* et *in cellulo*. Plus précisément, la présence de séries de cytosines et de régions riches en structures secondaires, basées sur des paires de bases Watson-Crick, risquent de séquestrer les guanines impliquées dans le G-quadruplexe empêchant sa formation. Ces observations soulèvent quelques questions. Quelle est la distance en 5' et en 3' à prendre en considération lorsque l'on veut prédire adéquatement la formation de G-quadruplexe? Une étude plus approfondie d'un nouveau candidat, un G-quadruplexe présent dans le 3'-UTR du gène *Tweety*, a permis de se rendre compte que 10 à 15 nucléotides en 5' et 3' du G-quadruplexe était probablement sous optimal (résultats non publiés). Quelle valeur est la plus informative à ce sujet? Les résultats préliminaires d'une analyse *in vitro* de versions plus longues pour chacun des G-quadruplexes présentés dans cette thèse, c'est-à-dire en tenant compte de 50 nucléotides en 5' et 3', a permis d'identifier et de suggérer une nouvelle valeur de prédiction de la formation de G-quadruplexe *in vitro* et *in cellulo* (résultats non publiés).

7.1 Le G-quadruplexe dans le 3'-UTR de l'ARNm du gène Tweety

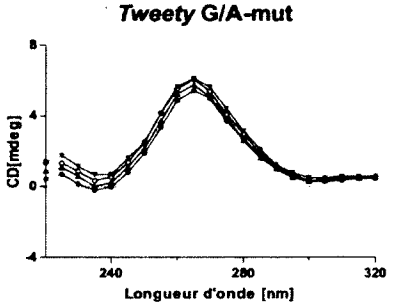
Le 3'-UTR du gène *Tweety* possède un G4P avec cinq séries de guanines (Discussion, Figure 10A). Cette séquence a été démontrée pour former une structure G-quadruplexe *in vitro* en présence de concentration physiologique de K^+ par les trois techniques utilisées dans les chapitres 2 et 3 (Discussion, Figure 10B-E). Différentes constructions avec un gène rapporteur ont été fabriquées pour diverses versions du 3'-UTR de *Tweety* (Discussion, Figure 10F). L'expression génique au niveau protéique de ces constructions a été étudiée et comparée avec celle des constructions de *LRP5* présentées au chapitre 3. Plusieurs éléments de la séquence primaire diffèrent pour chacun de ces deux 3'-UTR. Comme il a été

A *Tweety* 5' GGGAGUAGCUGAGGGGGCAGACUAGGGAGUAGGGCUGGCAGGGGAGGGGCAGACAGCCUCGC 3'

B



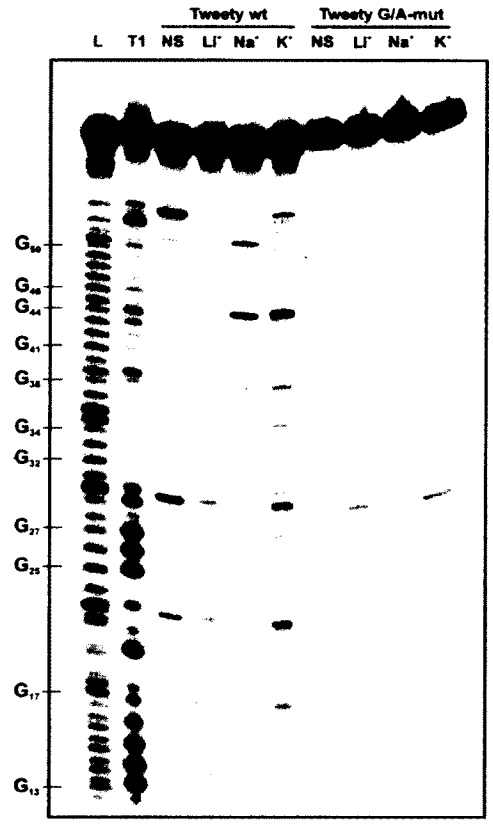
C



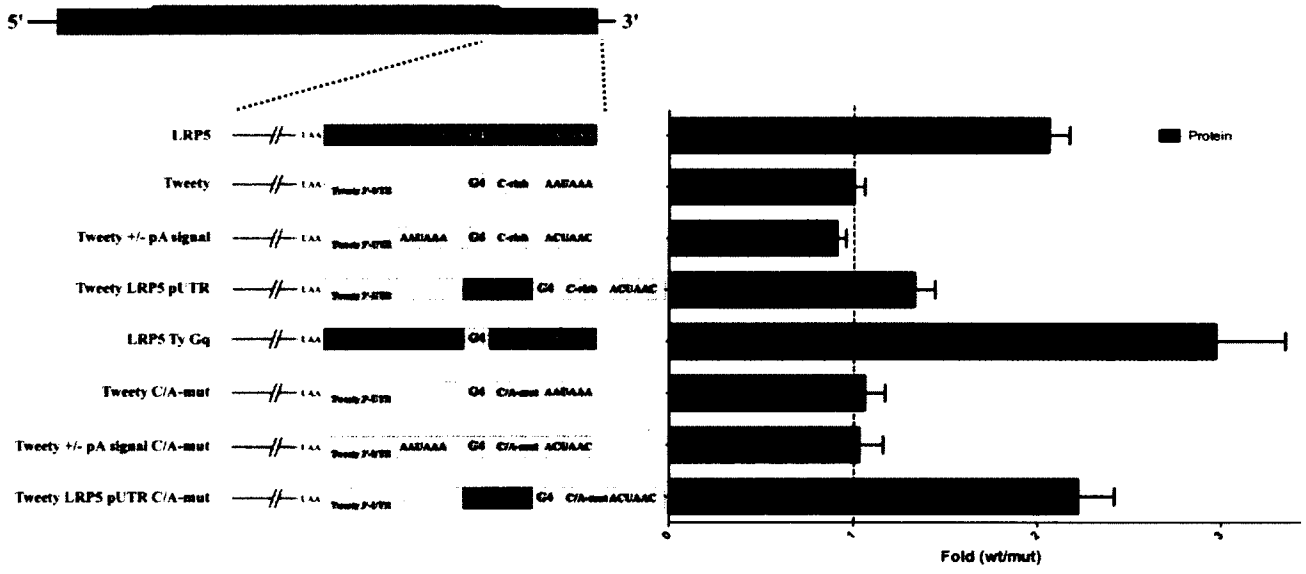
D

3' UTR		No salt	Li+	Na+	K+
Tweety	wt	48.4 ± 0.3	55.2 ± 2.1	53.4 ± 0.9	78.2 ± 1.8
	G/A-mut	52.5 ± 3.5	62.7 ± 0.7	62.4 ± 1.3	60.6 ± 0.6

E



F



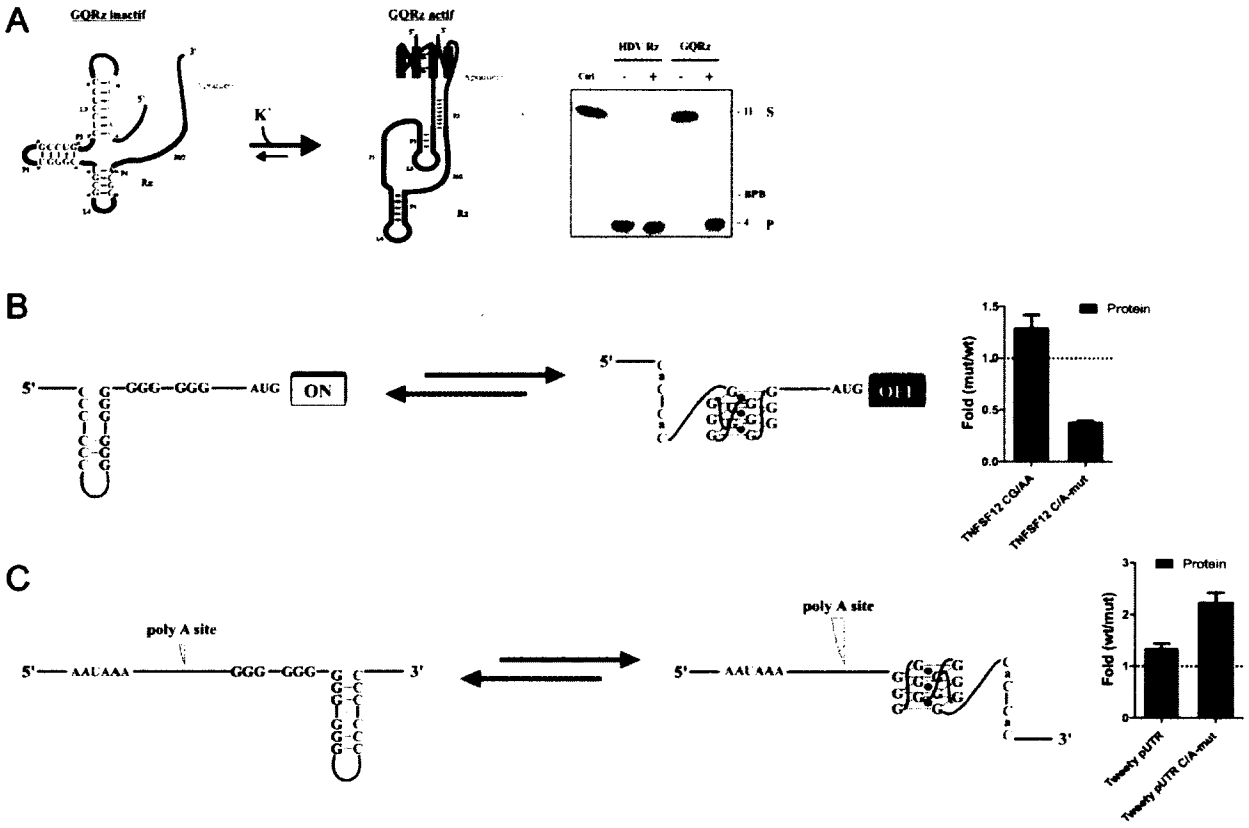
Discussion, Figure 10. Caractérisation du G-quadruplexe présent dans le 3'-UTR du gène *Tweety*.

(**A**) Séquence primaire du candidat *Tweety* utilisée pour l'analyse *in vitro*. La séquence correspondant au G4P est encadrée en jaune. Les guanines en minuscule représentent les guanines substituées en adénines dans la version G/A-mut. Les nucléotides soulignés correspondent aux positions devenant significativement plus accessibles en présence de K^+ *versus* en présence de Li^+ après quantification. (**B**,**C**) Spectres de dichroïsme circulaire pour *Tweety* en utilisant 4 µM d'ARN de la version wt (**B**) ou G/A-mut (**C**). Données recueillies soit en absence de sel (cercles pleins) ou en présence de 100 mM LiCl (triangles pleins), NaCl (cercles claires) ou KCl (triangles claires). (**D**) Valeurs de dénaturation thermique obtenues pour les versions wt et G/A-mut en absence de sel (No salt) ou en présence de 100 mM LiCl, NaCl et KCl. (**E**) Autoradiogramme typique d'un gel de cartographie par "in-line probing" des versions de *Tweety* (wt et G/A-mut) radiomarquées en 5' soit en absence de sel (NS) ou en présence de 100 mM de Li^+, de Na^+ ou de K^+. Les lignes L et T1 représentent respectivement une échelle moléculaire d'une hydrolyse alcaline et d'une ribonucléase T1 produite avec la version wt. La localisation de certaines guanines est représentée à la gauche du gel. (**F**) Histogramme des résultats de l'analyse de l'expression génique *in cellulo* réalisée par essai luciférase pour les différentes constructions utilisées. Des schémas représentant chacune des constructions utilisées sont présentés à la gauche de l'histogramme. Les parties en jaune correspondent à des séquences du 3'-UTR de *Tweety*. Les parties en bleu correspondent à des séquences du 3'-UTR de *LRP5*. L'axe des X identifie la différence de "*Fold*" qui correspond à la valeur obtenue pour la version wt du G-quadruplexe divisée par celle de la version G/A-mut pour chacune des constructions.

mentionné au chapitre 3, un signal de polyadénylation alternatif, en plus du canonique, est retrouvé 48 nucléotides en amont de la structure G-quadruplexe de *LRP5* (Chapitre 3, Figure 2D). En ce qui concerne le 3'-UTR de *Tweety*, seulement le signal de polyadénylation du site canonique est présent (Discussion, Figure 10F). Toutefois, il se caractérise par la présence d'une région riche en cytosines située à environ 20 nucléotides en aval de la structure G-quadruplexe de *Tweety* (Discussion, Figure 10F). Lorsque l'on compare l'expression génique en absence ou en présence du G-quadruplexe, une augmentation de l'expression est observée en présence de cette structure pour *LRP5*, alors qu'aucune différence n'est perceptible pour *Tweety* (Discussion, Figure 10F). La séquence du 3'-UTR de *Tweety* fût modifiée afin d'y insérer un site de polyadénylation alternatif sous la gouverne du G-quadruplexe. Pour se faire, deux stratégies furent utilisées. La

première comportait l'ajout d'un signal de polyadénylation 49 nucléotides en amont de la structure G-quadruplexe. La seconde consistait à insérer le signal de polyadénylation de *LRP5*, ainsi que les nucléotides situés entre le signal de polyadénylation et la structure G-quadruplexe, immédiatement en 5' du G-quadruplexe de *Tweety* (Discussion, Figure 10F). Cette dernière construction permet de conserver certains éléments en *cis* importants pour le processus de polyadénylation potentiellement retrouvés au niveau du site alternatif actif de *LRP5*. Dans les deux cas, le signal de polyadénylation du site canonique de *Tweety* a été muté afin de l'inactiver et de se rapprocher du contexte retrouvé pour *LRP5* (Discussion, Figure 10F). Aucune de ces constructions n'était significativement affectée par la présence ou l'absence du G-quadruplexe au niveau de l'expression génique (Discussion, Figure 10F). Afin de savoir si la structure G-quadruplexe de *Tweety* était belle et bien capable de stimuler la polyadénylation, une nouvelle construction fût réalisée où le G-quadruplexe de *LRP5* fût remplacé par celui de *Tweety* dans le 3'-UTR de *LRP5* (Discussion, Figure 10F). Effectivement, le G-quadruplexe de *Tweety* semble être apte à stimuler la polyadénylation une fois dans le contexte génique du 3'-UTR *LRP5* et, ce, à un niveau légèrement supérieur au G-quadruplexe *LRP5*. Intrigué par l'incapacité du G-quadruplexe *Tweety* à activer la polyadénylation dans son propre contexte génique, trois nouvelles constructions furent fabriquées. Elles étaient identiques aux trois constructions réalisées précédemment, à l'exception que leur région riche en cytosines a été mutée en substituant plusieurs d'entre elles pour des adénosines (Discussion, Figure 10F). En effet, une analyse de la structure secondaire de cette région du 3'-UTR suggérait que cette région riche en cytosines viendrait former une longue tige boucle stable séquestrant les guanines du G-quadruplexe *Tweety*. Aucune différence d'expression ne fût observée pour les deux premières constructions. Toutefois, celle contenant le signal de polyadénylation alternatif de *LRP5* démontra une stimulation de la polyadénylation par la présence du G-quadruplexe *Tweety* comparativement à la même construction sans mutation de la région riche en cytosines (Discussion, Figure 10F). En résumé, une étude approfondie du candidat *Tweety* a permis de

démontrer que la présence d'une région riche en cytosines pouvait également moduler la formation d'un G-quadruplexe présent dans un 3'-UTR et de modifier sa fonction et son impact sur l'expression génique *in cellulo*.



Discussion, Figure 11. Impact des séries de cytosines sur la formation des G-quadruplexes.

(**A**) Le mécanisme d'action du GQRz en mettant l'emphase sur l'interaction inhibitrice conférée par la présence de plusieurs cytosines. (**B**) Schéma de l'impact de la présence de séries de cytosines dans la capacité à réprimer la traduction du G-quadruplexe *TNFSF12*. La mutation de certaines cytosines permet la formation du G-quadruplexe et une répression de la traduction *in cellulo*. L'axe des Y représente le "*Fold*", c.-à-d. les valeurs obtenues pour les mutants identifiés par l'axe des X divisées par celles obtenues pour le wt. (**C**) Schéma de l'impact de la présence de séries de cytosines dans la capacité à activer la polyadénylation du G-quadruplexe *Tweety*. La mutation de certaines cytosines permet la formation du G-quadruplexe et une augmentation de l'expression génique reflétant probablement une meilleur polyadénylation *in cellulo*. L'axe des Y représente le "*Fold*", c.-à-d. les valeurs obtenues pour les versions G-quadruplexe wt divisées par celles obtenues pour les versions G-quadruplexe G/A-mut des constructions identifiées par l'axe des X.

En conclusion, l'importance de la séquence primaire du G-quadruplexe et des séquences adjacentes à celui-ci est notoire concernant la formation de G-quadruplexes. En effet, l'influence de ce contexte génique fût observée pour trois mécanismes et rôles complètement différents de la structure G-quadruplexe. Tout d'abord, une série de cytosines présente dans la région P3-L3 du GQRz fût directement impliquée dans le mécanisme d'action de cette nouvelle classe de ribozyme et modifia la formation du G-quadruplexe dans ce contexte (Chapitre 1 et Discussion, Figure 11A). Deuxièmement, la présence de séries de cytosines, dans trois des neuf candidats des G-quadruplexes dans les 5'-UTR, empêcha leur formation *in vitro* et leur capacité à réprimer la traduction *in cellulo* pour *TNFSF12*, qui fût étudié davantage (Chapitre 2 et Discussion, Figure 11B). Dans ce cas, la mutation de la région riche en cytosines mena à une augmentation de la répression de la traduction. Finalement, la présence d'une région riche en cytosines dans le 3'-UTR de *Tweety* empêcha cette structure de stimuler la polyadénylation lorsque tous éléments en *cis* étaient favorables à ce processus (Chapitre 3 et Discussion, Figure 11C). Ici, la mutation de plusieurs cytosines permettait au G-quadruplexe de finalement stimuler l'expression génique. Pour terminer, le contexte génique fût démontré pour être d'une importance indéniable en ce qui concerne l'étude de la formation et des différents rôles des G-quadruplexes d'ARN, autant *in vitro* que *in cellulo*. Tous les mécanismes rapportés dans cette thèse n'échappant pas à cette réglementation.

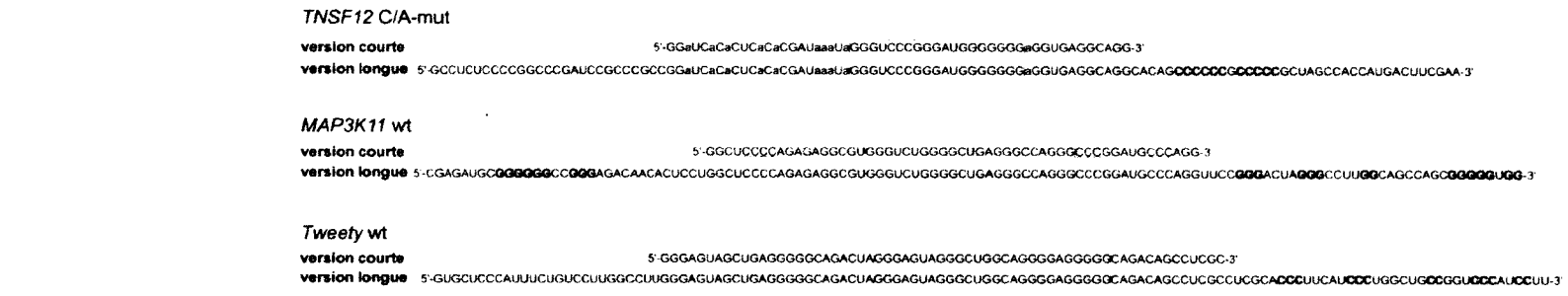
7.2 Comment tenir compte du contexte génique

Le contexte génique dans lequel une structure G-quadruplexe d'ARN se trouve est crucial à sa formation et à l'exercice de son ou ses rôles. Toutefois, comment passer de la théorie à la pratique? Comment approcher chaque G4P dans le transcriptome cellulaire en tenant compte de ce contexte génique? Selon moi, deux questions doivent être adressées et répondues afin d'atteindre cet objectif. En premier lieu, il faut déterminer la grandeur de la région à l'intérieur de laquelle le contexte génique maintient son impact sur la structure G-quadruplexe. En

second lieu, il faut identifier une valeur ou un paramètre, facile à calculer, donnant de l'information à savoir si une région donnée possède un contexte génique favorable ou non à la formation de G-quadruplexe.

Tout d'abord, en ce qui concerne la grandeur de la région à conserver et à considérer, il semble probable que 10 à 15 nucléotides en 5' et 3' du G-quadruplexe, bien que déjà informatif, ne soit pas suffisant. L'étude du candidat *Tweety* supporte bien cette déclaration. La séquence étudiée pour *Tweety* adoptait une structure G-quadruplexe stable *in vitro*, en gardant 13 nucléotides en 3' (Discussion, Figure 10A-E). Une fois dans le contexte de son 3'-UTR complet, il a été démontré qu'une région riche en cytosines située à partir de 20 nucléotides en 3' du G-quadruplexe était responsable d'une perte de fonction du G-quadruplexe *in cellulo* (Discussion, Figure 10F). Dans le but d'explorer ce phénomène, tous les G-quadruplexes présentés dans cette thèse (c.-à-d. les versions wt, G/A-mut, C/A-mut et CG/AA-mut des G-quadruplexes retrouvés dans les 5'- et 3'-UTR) ont été étudiés à nouveau, mais cette fois avec des versions plus longues considérant et conservant 50 nucléotides en 5' et 50 nucléotides en 3' du G-quadruplexe. Une comparaison entre la version précédente (maintenant appelée version courte) et cette nouvelle version (appelée version longue) est représentée à la Discussion, Figure 6 de la discussion pour le candidat *FXR1*. L'habileté de ces versions longues à adopter une structure G-quadruplexe *in vitro* a été testée par Mlle Rachel Jodoin avec la technique de "*in-line probing*" tel que mentionné à la section 3.2 de la discussion. Les résultats préliminaires obtenus pour les versions longues de chaque candidat furent comparés à ceux récoltés pour les versions courtes. À la lumière des résultats préliminaires obtenus et de façon sommaire, la formation de G-quadruplexe par la version courte et la version longue d'un même candidat semble généralement bien corréler. Toutefois, quelques différences ont été observées, par exemple pour les séquences *TNSF12* C/A-mut, *MAP3K11* wt et *Tweety* wt. Plus précisément, les séquences correspondant aux versions longues de *TNSF12* C/A-mut et *Tweety* wt ne semblent pas former de structure G-quadruplexe alors que leurs homologues courts démontrent l'inverse. Finalement, la version longue de *MAP3K11* wt semble adopter une structure G-quadruplexe,

mais pas sa version courte. En portant une attention plus particulière à la nature des séquences excédantes à la version courte de chacun de ces candidats, on peut constater la présence de plusieurs séries de cytosines pour *TNSF12* C/A-mut et *Tweety* wt et plusieurs séries de guanosines pour *MAP3K11* wt (Discussion, Figure 12). Sachant que des séries de cytosines peuvent empêcher la formation de G-quadruplexes, il n'est pas surprenant que ces versions longues ne puissent pas adopter cette structure *in vitro*. Pour ce qui est de la séquence *MAP3K11* wt, la présence de plusieurs séries de cytosines à l'intérieur de sa version courte a déjà été identifiée pour inhiber la formation de ce G-quadruplexe (Chapitre 2, Figure 2). Il est donc raisonnable d'imaginer que ces séries de guanosines supplémentaires puissent interagir avec les séries de cytosines, qui séquestraient initialement le G-quadruplexe, le libérant de cette structure secondaire inhibitrice et permettant la formation du G-quadruplexe. Encore une fois, ces trois exemples semblent démontrer que tenir compte de seulement 10 à 15 nucléotides en 5' et 3' du G4P n'est probablement pas suffisant dans plusieurs situations.



Discussion, Figure 12. Séquences primaires des candidats démontrant des différences de résultat entre leurs versions courte et longue *in vitro*.

Les séquences encadrées correspondent aux G4P et les séries de guanines ou de cytosines présentes dans les séquences excédant la version courte sont encadrées en jaune.

Bien qu'il semble être relativement facile de trouver une explication pour ces différences en prenant le temps de les regarder attentivement une à la fois, l'intérêt principal serait d'utiliser ces résultats afin de trouver une valeur offrant cette même

EBAG9

+40 +10 +10 +100 +100 +200 +350 +200 +40 +10

GGAGCCUCCGCCGGGCGGGCGGGGAGGGGGAGGGGCAGGUUUUGA

Nb G: 26 x 10 = 260 | Nb GG: 16 x 20 = 320 | Nb GGG: 9 x 30 = 270 | Nb GGGG: 4 x 40 = 160 | Nb GGGGG: 1 x 50 = 50 | cG score: **1060**

Nb C: 9 x 10 = 90 | Nb CC: 3 x 20 = 60 | Nb CCC: 0 x 30 = 0 | Nb CCCC: 0 x 40 = 0 | cC score: **150**

cG score / cC score: **7.1**

TNSF12

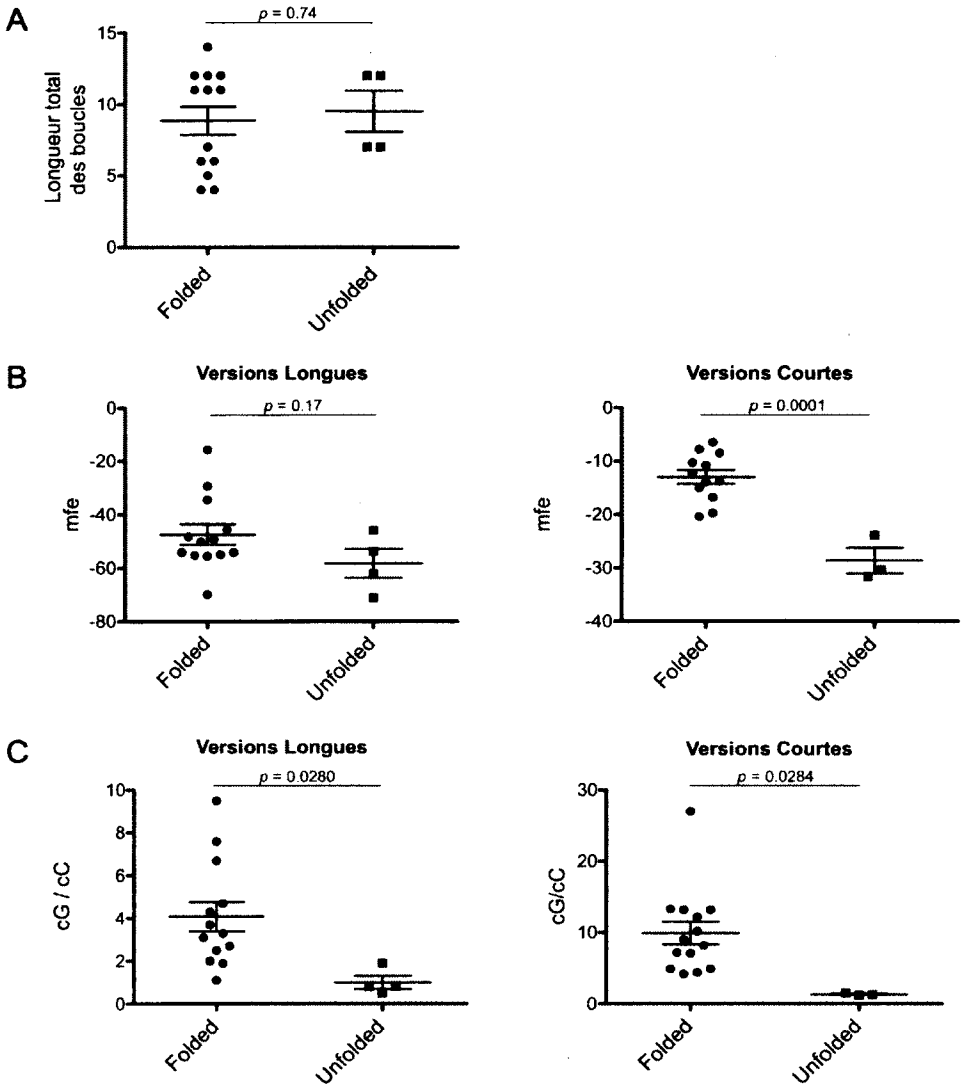
GGCUCCCCCUCCCCCGAUCCCUCGGGUCCCGGGAUGGGGGGGCGGUGAGGCAGG

cG score: **1220**
cC score: **940**
cG score / cC score: **1.3**

Discussion, Figure 13. Détails des calculs des cG, cC et cG/cC scores.

information. La découverte d'une valeur prédictive à la formation de G-quadruplexe pour une séquence donnée serait profitable. Dans cette optique, différentes valeurs ont été calculées pour chacune des séquences étudiées, les versions longues et courtes incluses. En premier lieu, la longueur pour chacune des boucles de chaque G4P a été notée et la longueur totale des boucles calculée. Deuxièmement, chaque séquence a été soumise au programme de prédiction de structure secondaire rnafold afin de trouver la structure prédite avec la plus faible valeur de "*minimum free energy*" (mfe) (Hofacker, 2003). Plus la valeur de mfe est basse, plus la séquence donnée est prédite pour former une structure secondaire stable basée sur les paires de bases Watson-Crick. Finalement, j'ai élaboré une toute nouvelle valeur appelée "*consecutive G score / consecutive C score*" (cG/cC). Le cG/cC est composé de deux valeurs le "*consecutive G score*" (cG score) et le "*consecutive C score*" (cC score). Les détails de chacune de ces valeurs sont présentés à la Discussion, Figure 13. Brièvement, le cG score prend en considération uniquement les guanosines d'une séquence donnée. Une valeur

numérique est attribuée à chaque guanosine de la séquence. Cette valeur est plus grande si la guanosine se retrouve dans une série de guanosines et elle augmente exponentiellement avec le nombre de guanosines dans la série. La valeur initiale pour une guanosine est de 10. Elle augmente par tranche de 10 pour chaque guanosine adjacente en 3' (Discussion, Figure 13). Pour une même position (guanosine), il peut y avoir une addition de plusieurs valeurs. Par exemple, si on se réfère au premier triplet de guanosines rencontré dans la séquence *EBAG9* (Discussion, Figure 13). La première guanosine de ce triplet (en position 13 de la séquence) se voit attribuer une première valeur de 10 pour une guanosine seule. Puisqu'une autre guanosine est présente en position 14 elle constitue également un doublet de guanosines et une valeur additionnelle de 20 est ajoutée au 10 précédent pour un total de 30 (Discussion, Figure 13). Finalement, une autre guanosine est présente en position 15 permettant de former un triplet de guanosines et d'attribuer une autre valeur additionnelle de 30. Elle s'ajoute au total précédent de 30, représentant un grand total de 60 (Discussion, Figure 13). La valeur numérique pour cette première guanine de ce triplet est donc de 60. Suivant la même logique, la valeur numérique attribuée à la seconde guanosine de cette série (position 14) est 30 et, finalement, une valeur de 10 pour la troisième guanosine (position 15). L'addition de ces trois valeurs correspond à la valeur attribuée pour le triplet en entier, soit 100 et restera la même pour chaque triplet rencontré dans toutes les séquences (Discussion, Figure 13). L'addition de ces valeurs pour toutes les guanosines d'une séquence donnée constitue son cG score. La même logique arithmétique est reprise en tenant compte uniquement des cytosines afin de calculer le cC score. Le ratio de ces valeurs compose finalement le cG/cC score (Discussion, Figure 13). Plus le cG/cC score est bas, plus la probabilité que les guanosines impliquées dans la structure G-quadruplexe soient séquestrées par des cytosines via des structures secondaires inhibitrices est grande.



Discussion, Figure 14. Analyse des différentes valeurs prédictives identifiées.

(**A**) Comparaison de la longueur total des boucles de chacun des candidats étudiés capables de former un G-quadruplexe ("*Folded*") ou non ("*Unfolded*") *in vitro*. (**B**) Comparaisons des valeurs de mfe pour les versions longues et les versions courtes de tous les candidats étudiés capables de former un G-quadruplexe ("*Folded*") ou non ("*Unfolded*") *in vitro*. (**C**) Comparaisons des valeurs de cG/cC score pour les versions longues et les versions courtes de tous les candidats étudiés capables de former un G-quadruplexe ("*Folded*") ou non ("*Unfolded*") *in vitro*. Les valeurs *P* ont été calculées par des tests de Student entre les groupes "*Folded*" et "*Unfolded*" en utilisant un intervalle de confiance de 95%.

Par la suite, la différence entre les valeurs des séquences, formant un G-quadruplexe *in vitro*, ont été comparées à celles n'adoptant pas cette structure et, ce, pour chacune des versions (courtes et longues). Tout d'abord, les données de la longueur totale des boucles ont été analysées. Cette valeur est la même pour les versions courtes et longues puisqu'elle se réfère au G4P uniquement. Étonnamment, aucune différence significative n'est observée entre les deux populations (Discussion, Figure 14A). Malgré les nombreuses publications réalisées sur l'étude de la longueur des boucles et la stabilité des G-quadruplexes, démontrant que de petites boucles augmentent copieusement la stabilité de la structure, ce critère ne semble pas être prédominant pour la prédiction de la formation de G-quadruplexe *in vitro* (Discussion, Figure 14A)(Guédin et al., 2010; Zhang et al., 2011). Le même exercice a été effectué pour les valeurs de mfe. Pour cette valeur, une différence significative a été observée entre ces deux populations pour les versions courtes alors que ce n'était pas le cas pour les versions longues, basée sur les résultats préliminaires récoltés (Discussion, Figure 14B). Les valeurs de mfe de l'environnement rapproché du G-quadruplexe semblent informatives sur la formation de G-quadrulpexes. Toutefois, en agrandissant la région d'intérêt, cette valeur semble perdre de son sens. Finalement, uniquement les valeurs de cG/cC score semblent être significativement différentes entre ces deux populations pour les versions courtes et longues (Discussion, Figure 14C). Le cG/cC score constituerait donc une très bonne valeur de prédiction de la formation de G-quadruplexe d'ARN. Il peut être calculé facilement par bioinformatique et utilisé pour analyser tous les G4P retrouvés dans le transcriptome.

Pour conclure, le contexte génique à l'intérieur duquel le G-quadruplexe se retrouve est manifestement l'un des critères les plus important, sinon le plus décisif, régissant la formation de cette structure. L'étude du candidat *Tweety* a démontré que ce contexte génique pouvait s'étendre à plus de 10 à15 nucléotides de par et d'autre du G-quadruplexe. À la lumière de mes résultats, la conservation de 50 nucléotides en 5' et 3' du G-quadruplexe et l'utilisation du cG/cC score comme valeur d'appréciation du contexte génique semble être une approche

prometteuse afin de prédire la formation de G-quadruplexe dans l'ARN. Cette approche pourrait être utilisée à la grandeur du transcriptome cellulaire où une valeur de cG/cC score serait accordée à chaque nucléotide du transcriptome en considérant une fenêtre de 100-125 nucléotides. Plus précisément, utiliser un système de pointage (ici le cG/cC score) au lieu d'un algorithme afin d'identifier des séquences G4P serait probablement plus efficace. En effet, de plus en plus d'exemples sont rapportés de structures G-quadruplexes biologiques ne remplissant pas les critères du fameux algorithme: G_x-N_{1-7}-G_x-N_{1-7}-G_x-N_{1-7}-G_x où x ≥ 3 et N correspond à n'importe lequel des quatre nucléotides (p. ex. les G-quadruplexes à longue boucle 2, CEB25, CEB1, *sc1* et le désoxyribozyme UVC1)(Chinnapen and Sen, 2004; Ribeyre et al., 2009; Phan et al., 2011; Amrane et al., 2012). Cet algorithme de prédiction, grandement utilisé jusqu'à aujourd'hui, a certainement apporté énormément au champ d'étude des G-quadruplexes *in cellulo*. Cependant, avec nos connaissances grandissantes concernant les règles régissant la formation de G-quadruplexes *in vitro* et *in cellulo*, l'instauration d'un système de pointage serait plus profitable et rigoureux à cette fin. De plus, il existe de plus en plus d'évidences que ces règles pour les G-quadruplexes d'ADN et les G-quadruplexes d'ARN sont différentes. Chacun de ces types de G-quadruplexes pourrait avoir son propre système de pointage spécifique, possédant ses propres règles et critères. Bien que davantage d'expériences et d'analyses doivent être réalisées à ce sujet, clairement le cG/cC score représente un bon point de départ dans l'élaboration d'un système de pointage pour la prédiction de la formation de G-quadruplexe dans une molécule d'ARN.

CONCLUSION

Cette thèse portait sur l'évaluation et l'étude des G-quadruplexes comme régulateurs de l'ARN. L'étude de cette structure dans la régulation de l'activité catalytique d'un ribozyme, dans la répression de la traduction des ARNm et dans la stimulation de la polyadénylation des pré-ARNm a apporté une multitude d'information à ce sujet. Elle a démontré toute la flexibilité que possède la structure G-quadruplexe afin d'interférer avec ces différents phénomènes liés à l'ARN. En effet, un G-quadruplexe peut mettre à profit sa séquence primaire ainsi que sa structure secondaire et tertiaire dans le but d'interagir avec son environnement. Cette étude a montré qu'une structure G-quadruplexe possède une communication étroite avec le milieu qui l'entoure. Le G-quadruplexe est capable de communiquer avec celui-ci et d'y entraîner certains changements spécifiques, à l'inverse son environnement est également capable d'affecter l'état de cette structure. C'est exactement l'une des qualités les plus importantes qu'un élément régulateur se doit de posséder. Les mécanismes d'action du G-quadruplexe comme régulateur de l'ARN semblent être variés. Sa formation ou non peut déterminer la structure globale d'une large portion d'ARN et modifier sa fonction. Le G-quadruplexe, avec sa grande stabilité, peut agir comme bloc stérique affectant certains processus. Bien que cette thèse se soit concentrée sur l'élément que représente le G-quadruplexe, il est pratiquement certain que cette structure interagit avec plusieurs facteurs protéiques afin de remplir plusieurs de ses rôles. L'identification de ces protéines est indubitablement d'une importance capitale afin de mieux comprendre l'impact des G-quadruplexes *in cellulo* et de mieux relier les données structurales des G-quadruplexes aux différents phénotypes observés. En terminant, il s'avère évident que les G-quadruplexes possèdent les propriétés et caractéristiques propices à faire de ces structures d'excellents éléments de régulation de l'ARN. D'ailleurs, ils semblent déjà avoir colonisés la majorité du génome et du transcriptome, où plusieurs rôles distincts leurs ont déjà été associés. Des milliers de G-quadruplexes y sont présents, il ne reste plus qu'à s'y attarder davantage!

REMERCIEMENTS

En tout premier lieu, je veux remercier mon directeur de recherche, le Dr Jean-Pierre Perreault, de m'avoir permis d'intégrer son laboratoire en 2005 pour mon deuxième stage du régime coopératif. L'intensité et l'entrain que Jean-Pierre apporte avec lui chaque jour sont certainement une source de motivation importante. Je le remercie particulièrement pour toute la confiance qu'il a su me donner. Il a su trouver le parfait équilibre entre encadrement et liberté de recherche. Malgré la nouveauté qu'elle apportait pour son groupe de recherche, il m'a permis de me laisser suivre ma passion pour les G-quadrulpexes. J'ai grandement apprécié le sentiment de collaboration présent entre nous deux. Jean-Pierre a à cœur la formation et la réussite de ses étudiants et il manque rarement une occasion de le mettre en action. Les nombreux conseils que Jean-Pierre m'a prodigués lors de toutes ces années ont eu un impact indéniable sur ma formation scientifique et influenceront positivement ma carrière scientifique à venir. Je lui en serai toujours reconnaissant.

Les personnes qui nous entourent à tous les jours dans le laboratoire sont certainement très importantes pendant nos études graduées. En ce sens, je me dois de remercier particulièrement, Dominique Lévesque, pour son support moral et technique. Ancien assistant de recherche du laboratoire, Dominique a été une source incommensurable de conseils techniques afin de faire débloquer plusieurs expériences sur la paillasse. Sa mémoire phénoménale et son expérience de laboratoire ont fait de cet homme un allié de premier plan à ma formation et au succès de mes études graduées. Je voudrais aussi remercier personnellement le Dr Jonathan Perreault, alias Jop. Jop m'a grandement aidé à attiser et entretenir ma passion scientifique, particulièrement au début de mes études graduées. Il m'a beaucoup aidé et encadré dans mon apprentissage de la programmation perl, qui a constitué le point de départ de tous les projets présentés dans cette thèse. Merci à Michel Lévesque pour sa bonne humeur, sa belle personnalité, sa joie de vivre et son esprit critique scientifique. Michel est un étudiant au doctorat ayant commencé

ses études graduées pratiquement en même temps que moi. Il a beaucoup contribué à l'ambiance exceptionnelle retrouvée dans le laboratoire. On ouvrira peut-être pas notre pizzéria Mike, mais je suis sur qu'on va bien se débrouiller pareil dans le futur.

Je veux également remercier personnellement deux autres étudiants du laboratoire, soit Samuel Rouleau et Rachel Jodoin. Samuel a permis de mettre mes idées concernant les G-quadruplexes à longue boucle 2 en action sur la paillasse en récoltant la majorité des résultats préliminaires discutés dans cette thèse. De son côté, Rachel a travaillé fort pour faire une analyse préliminaire par "*in-line probing*" de plusieurs des versions longues des G-quadruplexes présentées dans cette thèse. Les résultats préliminaires obtenus m'ont permis de développer le cG/cC score. Un gros merci à vous deux pour cette aide!

Merci également à tous les membres passés et présents du laboratoire. Mes études graduées n'auraient jamais été pareilles et aussi agréables sans vous. Merci aussi aux différents stagiaires qui se sont joints momentanément au laboratoire et qui ont contribué à diverses facettes de mes projets. Merci aux autres directeurs de recherche du département de biochimie, comme le Dr François Bachand et le Dr Martin Bisaillon pour toutes les discussions scientifiques qu'on a eues. Merci à mes collègues de travail des autres laboratoires pour l'ambiance hors pair présente au quotidien.

Je ne peux pas passer sous silence l'importance et l'impact majeur apporté par mes amis au cours de ces dernières années. La vie est remplie de hauts et de bas, tout comme la science d'ailleurs, et ton entourage est capital afin de partager les bons moments et de te supporter lors des moins bons. À ce sujet, je veux remercier Julie Motard qui, en plus d'avoir été une collègue de travail considérable, est devenue une amie exceptionnelle. Un merci tout spécial également à Alexis Dorais-Joncas et Mylène Brunelle pour toutes les discussions très intéressantes et

enrichissantes qu'on a eues sur les différents aspects de la vie. Merci aussi à ma famille pour leurs encouragements. Finalement, merci à mes parents et ma petite sœur préférée pour leur support dans la vie de tous les jours et leur intérêt apporté à mes recherches.

Je remercie les organismes subventionnaires, soient les Instituts de recherche en santé du Canada (IRSC) et le Fonds de recherche en santé du Québec (FRSQ), pour leur soutien financier au cours de mes études graduées.

Pour terminer, merci aux Drs Hervé Moine, François Bachand et Darel Hunting pour avoir accepté d'évaluer cette thèse.

RÉFÉRENCES

Amrane, S., Adrian, M., Heddi, B., Serero, A., Nicolas, A., Mergny, J.-L., and Phan, A.T. (2012). Formation of pearl-necklace monomorphic G-quadruplexes in the human CEB25 minisatellite. J. Am. Chem. Soc. *134*, 5807-5816.

Arnott, S., Chandrasekaran, R., and Marttila, C.M. (1974). Structures for polyinosinic acid and polyguanylic acid. Biochem. J. *141*, 537–543.

Arora, A., Dutkiewicz, M., Scaria, V., Hariharan, M., Maiti, S., and Kurreck, J. (2008). Inhibition of translation in living eukaryotic cells by an RNA G-quadruplex motif. RNA *14*, 1290–1296.

Bacolla, A., and Wells, R.D. (2009). Non-B DNA conformations as determinants of mutagenesis and human disease. Mol. Carcinog. *48*, 273–285.

Balkwill, G.D., Derecka, K., Garner, T.P., Hodgman, C., Flint, A.P.F., and Searle, M.S. (2009). Repression of translation of human estrogen receptor alpha by G-quadruplex formation. Biochemistry *48*, 11487–11495.

Baral, A., Kumar, P., Halder, R., Mani, P., Yadav, V.K., Singh, A., Das, S.K., and Chowdhury, S. (2012). Quadruplex-single nucleotide polymorphisms (Quad-SNP) influence gene expression difference among individuals. Nucleic Acids Res. *40*, 3800–3811.

Barák, I., Ricca, E., and Cutting, S.M. (2005). From fundamental studies of sporulation to applied spore research. Mol. Microbiol. *55*, 330–338.

Bardoni, B., Davidovic, L., Bensaid, M., and Khandjian, E.W. (2006). The fragile X syndrome: exploring its molecular basis and seeking a treatment. Expert Rev Mol Med *8*, 1–16.

Barrett, L.W., Fletcher, S., and Wilton, S.D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. Cell. Mol. Life Sci. *69*, 3613-3634.

Bartel, D.P. (2009). MicroRNAs: target recognition and regulatory functions. Cell *136*, 215–233.

Beaudoin, J.-D., and Perreault, J.-P. (2008). Potassium ions modulate a G-quadruplex-ribozyme's activity. RNA *14*, 1018–1025.

Beaudoin, J.-D., and Perreault, J.-P. (2010). 5'-UTR G-quadruplex structures acting as translational repressors. Nucleic Acids Res. *38*, 7022–7036.

Bechara, E., Davidovic, L., Melko, M., Bensaid, M., Tremblay, S., Grosgeorge, J., Khandjian, E.W., Lalli, E., and Bardoni, B. (2007). Fragile X related protein 1 isoforms differentially modulate the affinity of fragile X mental retardation protein for G-quartet RNA structure. Nucleic Acids Res. *35*, 299–306.

Bergeron, L.J., and Perreault, J.-P. (2005). Target-dependent on/off switch increases ribozyme fidelity. Nucleic Acids Res. *33*, 1240–1248.

Bergeron, L.J., Reymond, C., and Perreault, J.-P. (2005). Functional characterization of the SOFA delta ribozyme. RNA *11*, 1858–1868.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Stamatoyannopoulos, J.A., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Bonnal, S., Schaeffer, C., Créancier, L., Clamens, S., Moine, H., Prats, A.-C., and Vagner, S. (2003). A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons. J. Biol. Chem. *278*, 39330–39336.

Breaker, R.R. (2012). Riboswitches and the RNA world. Cold Spring Harb Perspect Biol *4*, doi: 10.1101/cshperspect.a003566.

Brown, V., Jin, P., Ceman, S., Darnell, J.C., O'Donnell, W.T., Tenenbaum, S.A., Jin, X., Feng, Y., Wilkinson, K.D., Keene, J.D., et al. (2001). Microarray identification of FMRP-associated brain mRNAs and altered mRNA translational profiles in fragile X syndrome. Cell *107*, 477–487.

Broxson, C., Beckett, J., and Tornaletti, S. (2011). Transcription arrest by a G quadruplex forming-trinucleotide repeat sequence from the human c-myb gene. Biochemistry *50*, 4162–4172.

Bryan, T.M., and Baumann, P. (2011). G-quadruplexes: from guanine gels to chemotherapeutics. Mol. Biotechnol. *49*, 198–208.

Bugaut, A., and Balasubramanian, S. (2012). 5'-UTR RNA G-quadruplexes: translation regulation and targeting. Nucleic Acids Res. *40*, 4727-4741.

Burge, S., Parkinson, G.N., Hazel, P., Todd, A.K., and Neidle, S. (2006). Quadruplex DNA: sequence, topology and structure. Nucleic Acids Res. *34*, 5402–5415.

Campbell, N.H., and Parkinson, G.N. (2007). Crystallographic studies of quadruplex nucleic acids. Methods *43*, 252–263.

Chantot, J.F., Sarocchi, M.T., and Guschlbauer, W. (1971). Physico-chemical properties of nucleosides. 4. Gel formation by quanosine and its analogues. Biochimie *53*, 347–354.

Chen, C.Y., and Shyu, A.B. (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends Biochem. Sci. *20*, 465–470.

Chinnapen, D.J.-F., and Sen, D. (2004). A deoxyribozyme that harnesses light to repair thymine dimers in DNA. Proc. Natl. Acad. Sci. U.S.a. *101*, 65–69.

Cho, H.-H., Cahill, C.M., Vanderburg, C.R., Scherzer, C.R., Wang, B., Huang, X., and Rogers, J.T. (2010). Selective translational control of the Alzheimer amyloid precursor protein transcript by iron regulatory protein-1. Journal of Biological Chemistry *285*, 31217–31232.

Cogoi, S., and Xodo, L.E. (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Res. *34*, 2536–2549.

Collie, G.W., and Parkinson, G.N. (2011). The application of DNA and RNA G-quadruplexes to therapeutic medicines. Chem Soc Rev. 40, 5867-5892.

Crick, F. (1970). Central dogma of molecular biology. Nature *227*, 561–563.

Dahan, O., Gingold, H., and Pilpel, Y. (2011). Regulatory mechanisms and networks couple the different phases of gene expression. Trends Genet. *27*, 316–322.

Dahm, R. (2008). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet. *122*, 565-581.

Dai, J., Dexheimer, T.S., Chen, D., Carver, M., Ambrus, A., Jones, R.A., and Yang, D. (2006). An intramolecular G-quadruplex structure with mixed parallel/antiparallel G-strands formed in the human BCL-2 promoter region in solution. J. Am. Chem. Soc. *128*, 1096–1098.

Darnell, J.C., Jensen, K.B., Jin, P., Brown, V., Warren, S.T., and Darnell, R.B. (2001). Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. Cell *107*, 489–499.

De Armond, R., Wood, S., Sun, D., Hurley, L.H., and Ebbinghaus, S.W. (2005). Evidence for the presence of a guanine quadruplex forming region within a polypurine tract of the hypoxia inducible factor 1alpha promoter. Biochemistry *44*, 16341–16350.

Decorsière, A., Cayrel, A., Vagner, S., and Millevoi, S. (2011). Essential role for the interaction between hnRNP H/F and a G quadruplex in maintaining p53 pre-mRNA 3'-end processing and function during DNA damage. Genes Dev. *25*, 220–225.

Di Giammartino, D.C., Nishida, K., and Manley, J.L. (2011). Mechanisms and consequences of alternative polyadenylation. Mol. Cell *43*, 853–866.

Dixon, I.M., Lopez, F., Tejera, A.M., Estève, J.-P., Blasco, M.A., Pratviel, G., and Meunier, B. (2007). A G-quadruplex ligand with 10000-fold selectivity over duplex DNA. J. Am. Chem. Soc. *129*, 1502–1503.

Doherty, E.A., and Doudna, J.A. (2000). Ribozyme structures and mechanisms. Annu. Rev. Biochem. *69*, 597–615.

Du, Z., Zhao, Y., and Li, N. (2009). Genome-wide colonization of gene regulatory elements by G4 DNA motifs. Nucleic Acids Res. *37*, 6784–6798.

Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., and Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev. *18*, 1618–1629.

Düchler, M. (2012). G-quadruplexes: targets and tools in anticancer drug design. J Drug Target. *20*, 389-400.

Eddy, J., and Maizels, N. (2006). Gene function correlates with potential for G4 DNA formation in the human genome. Nucleic Acids Res. *34*, 3887–3896.

Eddy, J., and Maizels, N. (2008). Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Res. *36*, 1321–1333.

Elkon, R., Zlotorynski, E., Zeller, K.I., and Agami, R. (2010). Major role for mRNA stability in shaping the kinetics of gene induction. BMC Genomics *11*, 259.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636–640.

Fabian, M.R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. Annu. Rev. Biochem. *79*, 351–379.

Fialcowitz, E.J., Brewer, B.Y., Keenan, B.P., and Wilson, G.M. (2005). A hairpin-like structure within an AU-rich mRNA-destabilizing element regulates trans-factor binding selectivity and mRNA decay kinetics. J. Biol. Chem. *280*, 22406–22417.

Fraser, C.S., and Doudna, J.A. (2007). Structural and mechanistic insights into hepatitis C viral translation initiation. Nat. Rev. Microbiol. *5*, 29–38.

Gellert, M., Lipsett, M.N., and Davies, D.R. (1962). Helix formation by guanylic acid. Proc. Natl. Acad. Sci. U.S.a. *48*, 2013–2018.

Gomez, D., Lemarteleur, T., Lacroix, L., Mailliet, P., Mergny, J.-L., and Riou, J.-F. (2004). Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing. Nucleic Acids Res. *32*, 371–379.

Guédin, A., Gros, J., Alberti, P., and Mergny, J.-L. (2010). How long is too long? Effects of loop size on G-quadruplex stability. Nucleic Acids Res. *38*, 7858–7868.

Gupta, G., Garcia, A.E., Guo, Q., Lu, M., and Kallenbach, N.R. (1993). Structure of a parallel-stranded tetramer of the Oxytricha telomeric DNA sequence dT4G4. Biochemistry *32*, 7098–7103.

Halder, K., Halder, R., and Chowdhury, S. (2009a). Genome-wide analysis predicts DNA structural motifs as nucleosome exclusion signals. Mol Biosyst *5*, 1703–1712.

Halder, K., Largy, E., Benzler, M., Teulade-Fichou, M.-P., and Hartig, J.S. (2011). Efficient suppression of gene expression by targeting 5'-UTR-based RNA quadruplexes with bisquinolinium compounds. Chembiochem *12*, 1663–1668.

Halder, K., Wieland, M., and Hartig, J.S. (2009b). Predictable suppression of gene expression by 5'-UTR-based RNA quadruplexes. Nucleic Acids Res. *37*, 6811–6817.

Halvorsen, M., Martin, J.S., Broadaway, S., and Laederach, A. (2010). Disease-associated mutations that alter the RNA structural ensemble. PLoS Genet. *6*, e1001074.

Henderson, E., Hardin, C.C., Walk, S.K., Tinoco, I., and Blackburn, E.H. (1987). Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. Cell *51*, 899–908.

Henkin, T.M. (2008). Riboswitch RNAs: using RNA to sense cellular metabolism. Genes Dev. *22*, 3383–3390.

Hershman, S.G., Chen, Q., Lee, J.Y., Kozak, M.L., Yue, P., Wang, L.-S., and Johnson, F.B. (2008). Genomic distribution and functional analyses of potential G-quadruplex-forming sequences in Saccharomyces cerevisiae. Nucleic Acids Res. *36*, 144–156.

Higgins, D., and Dworkin, J. (2012). Recent progress in Bacillus subtilis sporulation. FEMS Microbiol. Rev. *36*, 131–148.

Hofacker, I.L. (2003). Vienna RNA secondary structure server. Nucleic Acids Res. *31*, 3429–3431.

Huang, Y., Shen, X.J., Zou, Q., Wang, S.P., Tang, S.M., and Zhang, G.Z. (2011). Biological functions of microRNAs: a review. J. Physiol. Biochem. *67*, 129–139.

Huppert, J.L. (2007). Four-stranded DNA: cancer, gene regulation and drug development. Philos Transact a Math Phys Eng Sci *365*, 2969–2984.

Huppert, J.L. (2008). Four-stranded nucleic acids: structure, function and targeting of G-quadruplexes. Chem Soc Rev *37*, 1375–1384.

Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Res. *33*, 2908–2916.

Huppert, J.L., Bugaut, A., Kumari, S., and Balasubramanian, S. (2008). G-quadruplexes: the beginning and end of UTRs. Nucleic Acids Res. *36*, 6260–6268.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Jackson, R.J. (2005). Alternative mechanisms of initiating translation of mammalian mRNAs. Biochem. Soc. Trans *33*, 1231.

Ji, X., Sun, H., Zhou, H., Xiang, J., Tang, Y., and Zhao, C. (2011). Research progress of RNA quadruplex. Nucl Acid Ther *21*, 185–200.

Kedde, M., van Kouwenhove, M., Zwart, W., Oude Vrielink, J.A.F., Elkon, R., and Agami, R. (2010). A Pumilio-induced RNA structure switch in p27-3' UTR controls miR-221 and miR-222 accessibility. Nat. Cell Biol. *12*, 1014–1020.

Kempf, B., and Bremer, E. (1998). Uptake and synthesis of compatible solutes as microbial stress responses to high-osmolality environments. Arch. Microbiol. *170*, 319–330.

Kikin, O., D'Antonio, L., and Bagga, P.S. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. *34*, W676–W682.

Kim, J.N., and Breaker, R.R. (2008). Purine sensing by riboswitches. Biol. Cell *100*, 1–11.

Kim, M.-Y., Vankayalapati, H., Shin-Ya, K., Wierzba, K., and Hurley, L.H. (2002). Telomestatin, a potent telomerase inhibitor that interacts quite specifically with the human telomeric intramolecular g-quadruplex. J. Am. Chem. Soc. *124*, 2098–2099.

Kirkpatrick, L.L., McIlwain, K.A., and Nelson, D.L. (1999). Alternative splicing in the murine and human FXR1 genes. Genomics *59*, 193–202.

Kumari, S., Bugaut, A., Huppert, J.L., and Balasubramanian, S. (2007). An RNA G-quadruplex in the 5' UTR of the NRAS proto-oncogene modulates translation. Nat. Chem. Biol. *3*, 218–221.

Lacroix, L., Séosse, A., and Mergny, J.-L. (2011). Fluorescence-based duplex-quadruplex competition test to screen for telomerase RNA quadruplex ligands. Nucleic Acids Res. *39*, e21.

Laederach, A., Das, R., Vicens, Q., Pearlman, S.M., Brenowitz, M., Herschlag, D., and Altman, R.B. (2008). Semiautomated and rapid quantification of nucleic acid footprinting and structure mapping experiments. Nat Protoc *3*, 1395–1401.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Larson, E.D., Duquette, M.L., Cummings, W.J., Streiff, R.J., and Maizels, N. (2005). MutSalpha binds to and promotes synapsis of transcriptionally activated immunoglobulin switch regions. Curr. Biol. *15*, 470–474.

Lévesque, D., Choufani, S., and Perreault, J.-P. (2002). Delta ribozyme benefits from a good stability in vitro that becomes outstanding in vivo. RNA *8*, 464–477.

Lévesque, M.V., Rouleau, S.G., and Perreault, J.-P. (2011). Selection of the most potent specific on/off adaptor-hepatitis delta virus ribozymes for use in gene targeting. Nucl Acid Ther *21*, 241–252.

Licatalosi, D.D., and Darnell, R.B. (2010). RNA processing and its regulation: global insights into biological networks. Nat. Rev. Genet. *11*, 75–87.

Lipps, H.J., and Rhodes, D. (2009). G-quadruplex structures: in vivo evidence and function. Trends Cell Biol. *19*, 414–422.

Lopes, J., Piazza, A., Bermejo, R., Kriegsman, B., Colosio, A., Teulade-Fichou, M.-P., Foiani, M., and Nicolas, A. (2011). G-quadruplex-induced instability during leading-strand replication. Embo J. *30*, 4033-4046.

Lu, M., Guo, Q., and Kallenbach, N.R. (1992). Structure and stability of sodium and potassium complexes of dT4G4 and dT4G4T. Biochemistry *31*, 2455–2459.

Mani, P., Yadav, V.K., Das, S.K., and Chowdhury, S. (2009). Genome-wide analyses of recombination prone regions predict role of DNA structural motif in recombination. PLoS ONE *4*, e4399.

Marcel, V., Tran, P.L.T., Sagne, C., Martel-Planche, G., Vaslin, L., Teulade-Fichou, M.-P., Hall, J., Mergny, J.-L., Hainaut, P., and Van Dyck, E. (2011). G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms. Carcinogenesis *32*, 271–278.

McKee, A.E., and Silver, P.A. (2007). Systems perspectives on mRNA processing. Cell Res. *17*, 581–590.

Melko, M., and Bardoni, B. (2010). The role of G-quadruplex in RNA metabolism: Involvement of FMRP and FMR2P. Biochimie *92*, 919–926.

Menendez, C., Frees, S., and Bagga, P.S. (2012). QGRS-H Predictor: a web server for predicting homologous quadruplex forming G-rich sequence motifs in nucleotide sequences. Nucleic Acids Res. *40*, W96-W103.

Mergny, J.-L., and Lacroix, L. (2009). UV Melting of G-Quadruplexes. Curr Protoc Nucleic Acid Chem *Chapter 17*, Unit17.1.

Millevoi, S., Moine, H., and Vagner, S. (2012). G-quadruplexes in RNA biology. Wiley Interdiscip Rev RNA *3*, 495–507.

Mirkin, S.M. (2007). Expandable DNA repeats and human disease. Nature *447*, 932–940.

Monchaud, D., Allain, C., Bertrand, H., Smargiasso, N., Rosu, F., Gabelica, V., De Cian, A., Mergny, J.L., and Teulade-Fichou, M.P. (2008). Ligands playing musical chairs with G-quadruplex DNA: a rapid and simple displacement assay for identifying selective G-quadruplex binders. Biochimie *90*, 1207–1223.

Moore, M.J. (2005). From birth to death: the complex lives of eukaryotic mRNAs. Science *309*, 1514–1518.

Morris, M.J., and Basu, S. (2009). An unusually stable G-quadruplex within the 5'-UTR of the MT3 matrix metalloproteinase mRNA represses translation in eukaryotic cells. Biochemistry *48*, 5313–5319.

Nakken, S., Rognes, T., and Hovig, E. (2009). The disruptive positions in human G-quadruplex motifs are less polymorphic and more conserved than their neutral counterparts. Nucleic Acids Res. *37*, 5749–5756.

Neidle, S., and Balasubramanian, S. (2006) Quadruplex nucleic acids. *RSC Publishing*, Cambridge.

Neidle, S., and Parkinson, G.N. (2008). Quadruplex DNA crystal structures and drug design. Biochimie *90*, 1184–1196.

Paeschke, K., Capra, J.A., and Zakian, V.A. (2011). DNA replication through G-quadruplex motifs is promoted by the Saccharomyces cerevisiae Pif1 DNA helicase. Cell *145*, 678–691.

Paeschke, K., Juranek, S., Simonsson, T., Hempel, A., Rhodes, D., and Lipps, H.J. (2008). Telomerase recruitment by the telomere end binding protein-beta facilitates G-quadruplex DNA unfolding in ciliates. Nat Struct Mol Biol *15*, 598–604.

Paeschke, K., Simonsson, T., Postberg, J., Rhodes, D., and Lipps, H.J. (2005). Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. Nat Struct Mol Biol *12*, 847–854.

Phan, A.T., Kuryavyi, V., Darnell, J.C., Serganov, A., Majumdar, A., Ilin, S., Raslin, T., Polonskaia, A., Chen, C., Clain, D., et al. (2011). Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. Nat Struct Mol Biol *18*, 796–804.

Piazza, A., Boulé, J.-B., Lopes, J., Mingo, K., Largy, E., Teulade-Fichou, M.-P., and Nicolas, A. (2010). Genetic instability triggered by G-quadruplex interacting Phen-DC compounds in Saccharomyces cerevisiae. Nucleic Acids Res. *38*, 4337–4348.

Pichon, X., Wilson, L.A., Stoneley, M., Bastide, A., King, H.A., Somers, J., and Willis, A.E.E. (2012). RNA Binding Protein/RNA Element Interactions and the Control of Translation. Curr. Protein Pept. Sci. *13*, 294–304.

Qin, Y., and Hurley, L.H. (2008). Structures, folding patterns, and functions of intramolecular DNA G-quadruplexes found in eukaryotic promoter regions. Biochimie *90*, 1149–1171.

Rankin, S., Reszka, A.P., Huppert, J., Zloh, M., Parkinson, G.N., Todd, A.K., Ladame, S., Balasubramanian, S., and Neidle, S. (2005). Putative DNA quadruplex formation within the human c-kit oncogene. J. Am. Chem. Soc. *127*, 10584–10589.

Rawal, P., Kummarasetti, V.B.R., Ravindran, J., Kumar, N., Halder, K., Sharma, R., Mukerji, M., Das, S.K., and Chowdhury, S. (2006). Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. Genome Res. *16*, 644–655.

Regulski, E.E., and Breaker, R.R. (2008). In-line probing analysis of riboswitches. Methods Mol. Biol. *419*, 53–67.

Ribeyre, C., Lopes, J., Boulé, J.-B., Piazza, A., Guédin, A., Zakian, V.A., Mergny, J.-L., and Nicolas, A. (2009). The yeast Pif1 helicase prevents genomic instability caused by G-quadruplex-forming CEB1 sequences in vivo. PLoS Genet. *5*, e1000475.

Rinn, J.L., and Chang, H.Y. (2012). Genome Regulation by Long Noncoding RNAs. Annu. Rev. Biochem. *81*, 145–166.

Saccà, B., Lacroix, L., and Mergny, J.-L. (2005). The effect of chemical modifications on the thermal stability of different G-quadruplex-forming oligonucleotides. Nucleic Acids Res. *33*, 1182–1192.

Sana, J., Faltejskova, P., Svoboda, M., and Slaby, O. (2012). Novel classes of non-coding RNAs and cancer. J Transl Med *10*, 103.

Scaria, V., Hariharan, M., Arora, A., and Maiti, S. (2006). Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. Nucleic Acids Res. *34*, W683–W685.

Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H.J., and Plückthun, A. (2001). In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. Proc. Natl. Acad. Sci. U.S.a. *98*, 8572–8577.

Sen, D., and Gilbert, W. (1988). Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. Nature *334*, 364–366.

Shih, I.-H., and Been, M.D. (2002). Catalytic strategies of the hepatitis delta virus ribozymes. Annu. Rev. Biochem. *71*, 887–917.

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., and Hurley, L.H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc. Natl. Acad. Sci. U.S.a. *99*, 11593–11598.

Simonsson, T. (2001). G-quadruplex DNA structures--variations on a theme. Biol. Chem. *382*, 621–628.

Simonsson, T., Pecinka, P., and Kubista, M. (1998). DNA tetraplex formation in the control region of c-myc. Nucleic Acids Res. *26*, 1167–1172.

Sissi, C., Gatto, B., and Palumbo, M. (2011). The evolving world of protein-G-quadruplex recognition: A medicinal chemist's perspective. Biochimie *93*, 1219–1230.

Skourti-Stathaki, K., Proudfoot, N.J., and Gromak, N. (2011). Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. Mol. Cell *42*, 794–805.

Smith, J.S., Chen, Q., Yatsunyk, L.A., Nicoludis, J.M., Garcia, M.S., Kranaster, R., Balasubramanian, S., Monchaud, D., Teulade-Fichou, M.-P., Abramowitz, L., et al. (2011). Rudimentary G-quadruplex-based telomere capping in Saccharomyces cerevisiae. Nat Struct Mol Biol *18*, 478–485.

Solomon, E.I., and Lever, A.B.P. (1999) Inorganic electronic structure & spectro scopy, Volume 1 : Methodology. *Wiley-Interscience*.

Soukup, G.A., and Breaker, R.R. (1999). Relationship between internucleotide linkage geometry and the stability of RNA. RNA *5*, 1308–1325.

Subramanian, M., Rage, F., Tabet, R., Flatter, E., Mandel, J.-L., and Moine, H. (2011). G-quadruplex RNA structure as a signal for neurite mRNA targeting. EMBO Rep. *12*, 697–704.

Sun, D., Guo, K., Rusche, J.J., and Hurley, L.H. (2005). Facilitation of a structural transition in the polypurine/polypyrimidine tract within the proximal promoter

region of the human VEGF gene by the presence of potassium and G-quadruplex-interactive agents. Nucleic Acids Res. *33*, 6070–6080.

Sun, D., Thompson, B., Cathers, B.E., Salazar, M., Kerwin, S.M., Trent, J.O., Jenkins, T.C., Neidle, S., and Hurley, L.H. (1997). Inhibition of human telomerase by a G-quadruplex-interactive compound. J. Med. Chem. *40*, 2113–2116.

Sundquist, W.I., and Klug, A. (1989). Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops. Nature *342*, 825–829.

Teulade-Fichou, M., Lacroix, L., and Mergny, J. (2007). ScienceDirect - Methods : Fluorescence-based melting assays for studying quadruplex ligands. Methods. *42*, 183-195.

Tipson, R.S. (1957). Phoebus Aaron Theodor Levene, 1869-1940.

Todd, A.K., Johnston, M., and Neidle, S. (2005). Highly prevalent putative quadruplex sequence motifs in human DNA. Nucleic Acids Res. *33*, 2901–2907.

Tran, P.L.T., Mergny, J.-L., and Alberti, P. (2011). Stability of telomeric G-quadruplexes. Nucleic Acids Res. *39*, 3282–3294.

Vannier, J.-B., Pavicic-Kaltenbrunner, V., Petalcorin, M.I.R., Ding, H., and Boulton, S.J. (2012). RTEL1 Dismantles T Loops and Counteracts Telomeric G4-DNA to Maintain Telomere Integrity. Cell *149*, 795–806.

Venczel, E.A., and Sen, D. (1993). Parallel and antiparallel G-DNA structures from a complex telomeric sequence. Biochemistry *32*, 6220–6228.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Verma, A., Halder, K., Halder, R., Yadav, V.K., Rawal, P., Thakur, R.K., Mohd, F., Sharma, A., and Chowdhury, S. (2008). Genome-wide computational and expression analyses reveal G-quadruplex DNA motifs as conserved cis-regulatory elements in human and related species. J. Med. Chem. *51*, 5641–5649.

Verma, A., Yadav, V.K., Basundra, R., Kumar, A., and Chowdhury, S. (2009). Evidence of genome-wide G4 DNA-mediated gene expression in human cancer cells. Nucleic Acids Res. *37*, 4194–4204.

Wan, Y., Kertesz, M., Spitale, R.C., Segal, E., and Chang, H.Y. (2011). Understanding the transcriptome through RNA structure. Nat. Rev. Genet. *12*, 641–655.

Wanrooij, P.H., Uhler, J.P., Shi, Y., Westerlund, F., Falkenberg, M., and Gustafsson, C.M. (2012). A hybrid G-quadruplex structure formed between RNA and DNA explains the extraordinary stability of the mitochondrial R-loop. Nucleic Acids Res. *40*, 10334-10344.

Wanrooij, P.H., Uhler, J.P., Simonsson, T., Falkenberg, M., and Gustafsson, C.M. (2010). G-quadruplex structures in RNA stimulate mitochondrial transcription termination and primer formation. Proc. Natl. Acad. Sci. U.S.a. *107*, 16072–16077.

Watson, J.D., and Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Webba da Silva, M. (2007). NMR methods for studying quadruplex nucleic acids. Methods *43*, 264–277.

Wells, R.D. (2007). Non-B DNA conformations, mutagenesis and disease. Trends Biochem. Sci. *32*, 271–278.

Westhof, E., and Romby, P. (2010). The RNA structurome: high-throughput probing. Nat Meth *7*, 965–967.

Wieland, M., and Hartig, J.S. (2007). RNA quadruplex-based modulation of gene expression. Chem. Biol. *14*, 757–763.

Williamson, J.R., Raghuraman, M.K., and Cech, T.R. (1989). Monovalent cation-induced structure of telomeric DNA: the G-quartet model. Cell *59*, 871–880.

Winter, J., and Diederichs, S. (2011). MicroRNA biogenesis and cancer. Methods Mol. Biol. *676*, 3–22.

Winter, J., Jung, S., Keller, S., Gregory, R.I., and Diederichs, S. (2009). Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat. Cell Biol. *11*, 228–234.

Wong, A., and Wu, G. (2003). Selective binding of monovalent cations to the stacking G-quartet structure formed by guanosine 5'-monophosphate: a solid-state NMR study. J. Am. Chem. Soc. *125*, 13895–13905.

Wong, T.N., and Pan, T. (2009). RNA folding during transcription: protocols and studies. Meth. Enzymol. *468*, 167–193.

Wongsurawat, T., Jenjaroenpun, P., Kwoh, C.K., and Kuznetsov, V. (2012). Quantitative model of R-loop forming structures reveals a novel level of RNA-DNA interactome complexity. Nucleic Acids Res. *40*, e16.

Xu, Y. (2011). Chemistry in human telomere biology: structure, function and targeting of telomere DNA/RNA. Chem Soc Rev *40*, 2719–2740.

Zhang, A.Y.Q., Bugaut, A., and Balasubramanian, S. (2011). A Sequence-Independent Analysis of the Loop Length Dependence of Intramolecular RNA G-Quadruplex Stability and Topology. Biochemistry. *50*, 7251-7258.

Zimmerman, S.B., Cohen, G.H., and DAVIES, D.R. (1975). X-ray fiber diffraction and model-building study of polyguanylic acid and polyinosinic acid. J. Mol. Biol. *92*, 181–192.

ANNEXES

1. Tous les Datasets peuvent être retrouvés dans le fichier .zip

2. Articles additionnels

2.1 ARTICLE: Modulating RNA structure and catalysis: lessons from small cleaving ribozymes.

2.2 ARTICLE: A novel structural rearrangement of hepatitis delta virus antigenomic ribozyme.

2.3 ARTICLE: In vitro selection and characterization of RNA aptamers binding thyroxine hormone.

Cell. Mol. Life Sci. (2009) 66:3937–3950
DOI 10.1007/s00018-009-0124-1

Cellular and Molecular Life Sciences

Modulating RNA structure and catalysis: lessons from small cleaving ribozymes

Cedric Reymond · Jean-Denis Beaudoin · Jean-Pierre Perreault

Received: 24 April 2009 / Revised: 30 July 2009 / Accepted: 31 July 2009 / Published online: 30 August 2009

Abstract RNA is a key molecule in life, and comprehending its structure/function relationships is a crucial step towards a more complete understanding of molecular biology. Even though most of the information required for their correct folding is contained in their primary sequences, we are as yet unable to accurately predict both the folding pathways and active tertiary structures of RNA species. Ribozymes are interesting molecules to study when addressing these questions because any modifications in their structures are often reflected in their catalytic properties. The recent progress in the study of the structures, the folding pathways and the modulation of the small ribozymes derived from natural, self-cleaving, RNA motifs have significantly contributed to today's knowledge in the field.

Keywords Catalytic RNA · Modulation · RNA folding · Riboswitch · Ribozyme · Structure

Introduction

RNA is a key molecule in life, and the understanding of its structure/function relationships is crucial in molecular biology and therefore has important implications in terms of human health [1]. However, we are as yet unable to accurately predict either the folding pathway or the active tertiary structure of RNA molecules from their primary sequences. Clearly, these two fundamental questions need to be addressed. A long-term goal is to be able to interpret an RNA sequence in terms of its functionally folded three-dimensional form. RNA molecules possess a hierarchical structure: the primary sequence determines the formation of the secondary structure elements, and the secondary structure in turn determines the tertiary folding [2–4]. RNA molecules fold sequentially from 5′ to 3′ using recurring stable submotifs [5, 6], and the folding intermediates tend to become increasingly stable as the tertiary interaction network progresses [7–9]. RNA can fold into multiple structures; however, only a single structure is usually functional. In order to fold correctly, the RNA must avoid the problem of folding into alternative, non-functional structures and kinetic traps [10–13].

Ribozymes are interesting molecules to study when addressing these RNA questions because modifications in their structures are reflected in their catalytic properties. More precisely, small ribozymes are suitable for this task since considerable progress has been made in the determination of their structures [14, 15], and a versatile toolbox is available for their study. Here we present a review of the recent progress in the study of the tertiary structures, the folding pathways and the modulation of the small ribozymes derived from natural, self-cleaving, RNA motifs.

Small self-cleaving ribozymes

This group of natural ribozymes includes five RNA species that have been derived from self-cleaving sequences ranging from ~40 to 200 nucleotides in length and possessing various secondary structures. One subgroup includes three self-catalytic RNA motifs that are part of species belonging to the brotherhood of small, circular, self-replicating RNAs and are essential components of the

C. Reymond · J.-D. Beaudoin · J.-P. Perreault (✉)
RNA Group/Groupe ARN, Département de biochimie,
Faculté de médecine et des sciences de la santé,
Université de Sherbrooke, Sherbrooke, QC J1H 5N4, Canada
e-mail: Jean-Pierre.Perreault@usherbrooke.ca

rolling circle replication mechanism of these infectious RNAs. According to this mechanism, the circular monomer strands are replicated into linear multimeric strands of complementary polarity that then are self-cleaved, on the basis of the RNA motifs, into monomers that, following circularization, participate in the next round of replication [16]. More specifically, the hammerhead and hairpin RNA motifs were obtained from viroids and viroid-like satellite RNAs of plant origin [17–22], while the HDV self-cleaving RNA motif was obtained from the hepatitis delta virus (HDV) that infects humans [23–25]. The hammerhead motif has also been detected in eukaryote satellite DNA transcripts obtained from newt, schistosome and cricket [26–28], while the HDV motif has also been detected in the human genome [29].

The members of the second subgroup of natural self-cleaving RNA motifs are found within the bodies of larger transcripts. For example, the VS self-cleaving motif was identified by accident in the Varkud satellite RNA, an abundant transcript from the circular mitochondrial Varkud satellite DNA found in numerous *Neurospora* species and in other simple eukaryotes [30]. In common with the above ribozymes, this self-catalytic motif, in vivo, processes a multimeric RNA into monomers [31]. Another example is the *glmS* self-cleaving motif that was identified, by computer analysis of the 5′ untranslated region (5′UTR) of the glucosamine-6-phosphate synthase (glmS) mRNA of certain gram-positive bacteria, to be a putative riboswitch [32]. In fact, upon the binding of glucosamine-6-phosphate (GlcN6P), the *glmS* motif has been shown to undergo a site-specific self-cleavage reaction making it the first, and as yet the only, natural allosteric ribozyme [33]. This cleavage dramatically decreases the half-life of glmS mRNA, and thus this ribozyme acts as a regulatory element controlling the glucosamine-6-phosphate synthase level [33].

All of these self-cleaving RNA motifs catalyze the cleavage of the RNA phosphodiester backbone through a transesterification reaction involving the attack of the vicinal 2′-hydroxyl group (2′-OH) on the scissile phosphate, yielding a 2′-3′-cyclic phosphate and a 5′-hydroxyl termini as products [34]. The mechanism involved is a bimolecular nucleophilic substitution (S_N2) that is general acid–base catalyzed and requires in-line geometry between the 2′ oxygen, the phosphate and the 5′ oxygen. The catalytic strategies and active site organization at the atomic level for each of these ribozymes have been extensively characterized and do not fall under the scope of this work since these aspects have been recently been reviewed [1, 15]. Finally, the presence of divalent metal cations is important in RNA folding in order to obtain well-defined structures, as well as in the catalysis, by functioning either as a general acid or as a general base when coordinated with water. These self-cleaving RNA motifs are metalloenzymes under physiological conditions, but are also active in the absence of divalent metals under certain conditions such as in the presence of high monovalent metal ion concentrations or at low pH levels [35, 36].

HDV ribozyme

The self-cleaving RNA motif derived from HDV varies from 85 to 95 nucleotides in length depending on the sequence variant and the polarity. It has been possible to separate this self-catalytic sequence into two molecules, thereby creating a *trans*-acting system where one molecule, the ribozyme, possesses the catalytic properties required to cleave multiple copies of the other molecule, the substrate. This separation can be achieved in several ways, the most frequent being the removal of the junction between stems I and II (Fig. 1a) [25]. The development of *trans*-acting ribozymes, for HDV as well as all other self-catalytic RNA motifs, has permitted experiments leading to the identification of the important structural features of these motifs [24, 25, 37]. For example, direct mutagenesis experiments led to the identification of the crucial nucleotides in the catalytic core, as well as identifying essential base-paired positions [for examples see Refs. 37, 38]. It should be noted, however, that an unbiased in vitro selection revealed that, with the exception of the catalytic nucleotide (position 76), none of the bases were absolutely required for cleavage [39]. This work clearly revealed that the ribozyme supports more variability than was originally thought to be the case based on the isolation of natural sequences. In addition, it also revealed that higher cleavage levels were observed for sequences closely related to the natural ones, suggesting that the combination of all of the structural features is important for optimal cleavage activity.

According to the experimentally well-supported pseudoknot model, the HDV ribozyme is composed of one stem (I or P1), one pseudoknot (II or P2), two stem-loops (stems III or P3 and IV or P4) and three single-stranded junctions (I/II, I/IV and IV/II). The structure depicted in Fig. 1a is derived from the HDV of antigenomic polarity. Previous nomenclature of the various regions and the harmonized one using roman numbers for the stems are indicated. It is composed of a 57-nucleotide ribozyme and an 11-nucleotide substrate. By using this design, the stem I is composed of one Wobble base pair followed by six Watson–Crick base pairs formed between the ribozyme and the substrate and becomes the recognition domain. Both the junction I/IV and the loop III are single-stranded in the initial folding steps, but are eventually involved in the formation of a second pseudoknot (I.I) [37, 40, 41]. This structure requires the presence of divalent cations

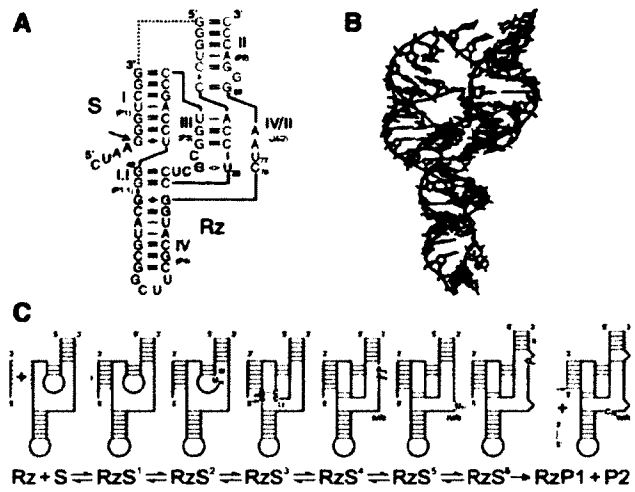


Fig. 1 Structure and folding pathway of HDV ribozyme. The ribozyme and substrate are denoted by *Rz* and *S*, respectively. The cleavage site is indicated by an *arrow*. **a** Secondary structure of a shortened *trans*-acting version of the wild-type antigenomic HDV ribozyme according to the notation of Leontis and Westhof [145]. The harmonized nomenclature using roman numbering for the stems is indicated as well as the previous one in *parenthesis*. The *dotted line* represents the junction that has been removed in order to generate a *trans*-acting version. **b** Crystal structure of the genomic version of HDV ribozyme [46]. The *colors* of the various domains are the same in both panels A and B. **c** Proposed folding pathway of HDV ribozyme according to Reymond et al. [8]. The substrate is shown in *green* and the products in *red*. Important nucleotides are indicated in the various intermediates

such as calcium or magnesium in order to be formed [42]. Both X-ray diffraction analysis and nuclear magnetic resonance studies have provided high-resolution tertiary structures of the genomic HDV ribozyme [40, 43, 44]. Overall, these approaches have shown that the global shape of the ribozyme is dominated by two coaxial helices formed by the stacking of the stems I–I.I–IV and the stems II–III (Fig. 1b). The stems II and IV have a structural role and are located above and below the catalytic core, which is formed by a network of interactions located at the interface between the two helical stacks. This network relies heavily on the nucleotides' identities, as well as on several of the 2′-OH groups of the ribose moieties [38, 45].

The crystal structure of the genomic HDV ribozyme also revealed that this catalytic RNA adopts a highly ordered structure [40, 43, 46], in agreement with the previously reported unusual properties of this motif. For example, the self-catalytic motif retains activity at temperatures as high as 80°C and in solutions containing up to 5 M urea [47, 48]. Moreover, it was shown that for some *trans*-acting HDV ribozyme variants the cleavage level was near 100%, suggesting that HDV ribozyme forms a homogeneous population of structures and is not prone to alternative structure formation [8]. This important characteristic has been fully exploited in order to decipher the folding pathway of the HDV ribozyme, by far the most complex folding pathway elucidated to date for a small ribozyme. Extensive work has been performed using a number of approaches, including in vitro selection and photo-induced cross linking, in order to identify a specific set of mutants able to halt the folding pathway at each stable intermediate [39, 49]. Briefly, after the recognition of the substrate by the ribozyme leading to the formation of the stem I, five subsequent conformational steps are required in order to form the catalytically active structure (Fig. 1c). The first conformational transition is the docking of the stem I within the catalytic core. A specific interaction between the substrate in the middle of the stem I, which is slightly unfolded at the position +4 from the cleavage site, and both the C22 and the U23 residues of the loop III are most likely responsible for this docking [50]. This conclusion is supported by the fact that replacing a weak base pair, such as UA, in the middle of the stem I by a stronger base pair, such as GC, is detrimental to the ribozyme's activity [51]. The substitution of an A for a U in position 23 blocks stem I docking to the catalytic core [50]. The second conformational transition is the formation of the pseudoknot I.I, which includes two GC base pairs [37, 41]. Any mutation of the pseudoknot I.I reducing either its stability, or the coaxial stacking, is detrimental to the cleavage activity [52]. The third conformational step is the formation of the A-minor motif between the two consecutive adenosines of the junction IV/II and the minor groove of the stem III, more specifically with the two GC base pairs of the latter [40]. This is a key interaction for the positioning of the junction IV/II, which is otherwise quite flexible. Moreover, this conformational transition initiates the positioning of the catalytic cytosine (C76), which is also located within the junction IV/II. The replacement of the adenosines in positions 78 and 79 by uridines prevents any further progression along the folding pathway [8, 40]. The fourth and fifth conformational steps are the formation of the trefoil motif and a base pair switch at the bottom of the stem II, respectively [40, 53, 54]. These two steps most likely occur simultaneously, although an order has been suggested based on physico-chemical analyses [8]. The trefoil turn is a particular motif identified by X-ray crystallography involving the catalytically active cytosine and its two flanking nucleotides [40]. This motif is induced by the formation of the A-minor motif and is proposed to contribute relaxing the phosphodiester backbone of the junction IV/II, thereby positioning the cytosine deep inside the catalytic centre [54]. Altering the trefoil turn motif by deleting the uridine in position 77 results in a ribozyme completely deprived of catalytic activity [53]. The other conformation transition, the base pair switch, takes place at the bottom of the stem II. It consists of switching the

C19–G81 base pair of the stem II to a new C19–G80 base pair and also bulging out the G81 residue [53]. The replacement of the guanosine in position 80 by a cytosine prevents the adoption of the ribozyme–substrate transition complex. This base pair switch is also induced by the formation of the A-minor motif. Upon completion of the base pair switch, the geometry at the catalytic site is correct, and the chemical reaction occurs. The nucleotide involved in this chemical step is the catalytically active cytosine C76, which plays the role of a general base [55, 56]. Any mutation of this cytosine results in the complete loss of cleavage activity. When the transesterification reaction occurs, it is simultaneously accompanied by the products' release. However, under certain conditions, it has been shown that the 3′-product might remain associated with the ribozyme, thereby causing product inhibition.

In summary, each of these conformational steps is separated by the formation of specific interactions. Several mutants have been designed in order to block the folding pathway at stable intermediates by disrupting these interactions. These mutants have been used in an isothermal titration calorimetry study that yielded the complete thermodynamic characterization of the HDV folding pathway [8]. This folding pathway is enthalpy driven, and the formation of additional interactions, such as the stacking in the coaxial helices, is responsible for the stability of this complex structure. The formation of the pseudoknot I.I has been shown to be the limiting step in the molecular mechanism of the HDV ribozyme [8, 46]. At the beginning of the folding pathway, the loop III and junction I/IV are located relatively far from each other and have a significant amount of freedom. The formation of the pseudoknot I.I is required in order to bring together these two single-stranded regions, thereby trapping the substrate within the catalytic core [49]. This requires an important entropic loss, and the driving force of this step is most likely the formation of two GC base pairs and the coaxial stacking of the helices I–I.I–IV.

Three different strategies have been used to modify activity of the HDV ribozyme. The first one is based on the fact that, due to its self-cleaving origin, the HDV ribozyme suffers from a lack of substrate specificity when used in *trans* as a molecular tool. Its substrate specificity depends on the formation of stem I that contains only seven base pairs (Fig. 1a), while a total of 15–16 base pairs has been estimated to be required in order to ensure the targeting of a unique RNA species from the human transcriptome [57]. This is the reason why the past interest in this ribozyme was only moderate. In order to overcome this hurdle a module named the SOFA (Specific On/ofF Adaptor) was engineered for the HDV ribozyme (Fig. 2a) [58]. The SOFA adaptor switches the cleavage activity from "off" to "on" state solely in the presence of its cognate substrate.

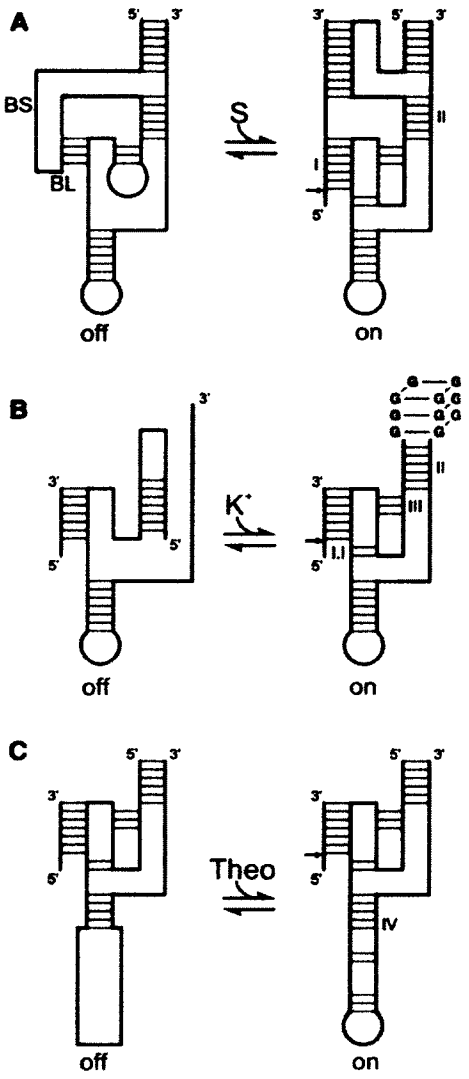


Fig. 2 Various modulated versions of the HDV ribozyme. The inactive and active states are indicated by "*off*" and "*on*", respectively. The substrate is shown in *green*, and the portion of the ribozyme responsible for the modulation in *blue*. **a** The SOFA module [58]. *BL* and *BS* indicate blocker and biosensor, respectively. **b** The G-quartzyme [64]. **c** The theophylline aptamer attached by a communication module [65]

This modulation acts at two different levels. The substrate binding site forms a short duplex with an inserted sequence element (the blocker); this increases the energetic barrier for non-specific base-pairing interactions with the substrate binding site, thus reducing the potential for off-target cleavages. A second inserted sequence element (the biosensor) extends base-pairing with the substrate to favor binding of the genuine substrate, and formation of this duplex concomitantly results in disruption of the short duplex involving the blocker sequence. Analysis of the cleavage activity using a large collection of substrates and

SOFA-ribozyme mutants provided evidence as to the roles of each domain and gave hints for design optimization [59]. The proof-of-concept of this adaptor was demonstrated both in vitro and in vivo using several SOFA-ribozymes that cleaved various RNA transcripts [58, 60, 61]. Several alternative versions of the SOFA module have been reported for the HDV ribozyme, and it has been adapted to various ribozymes including both the hammerhead and the hairpin ribozymes [62].

Two other ways by which the HDV ribozyme's activity can be modified involve the use of allosteric modules adapted to the HDV ribozyme. The first involves using an aptamer known to fold into an unstable hairpin structure in the absence of potassium and into a G-quadruplex structure upon the addition of potassium [63]. A rationally designed HDV ribozyme in which most of the stem II was replaced by this aptamer lost its cleavage activity in the absence of potassium (Fig. 2b) [64]. Further structural characterization led to the proposal that the 5′-strand of the aptamer was interacting with loop III, inactivating the HDV ribozyme. The addition of potassium promotes the formation of the G-quadruplex structure involving both strands of the aptamer, releasing the catalytic core of the ribozyme and stabilizing it by the formation of the stem II. The resulting RNA species represents a new class of ribozyme that exhibits a monovalent cation-dependent activity. The second method involves replacing most of the stem IV by a randomized region followed by a theophylline aptamer previously developed by in vitro selection (Fig. 2c) [65]. A negative selection was performed so as to remove all the sequences that cleave in the absence of theophylline. A positive selection was then performed in order to enrich the pool of sequences that are active in the presence of theophylline, thereby selecting a theophylline-dependent allosteric HDV ribozyme.

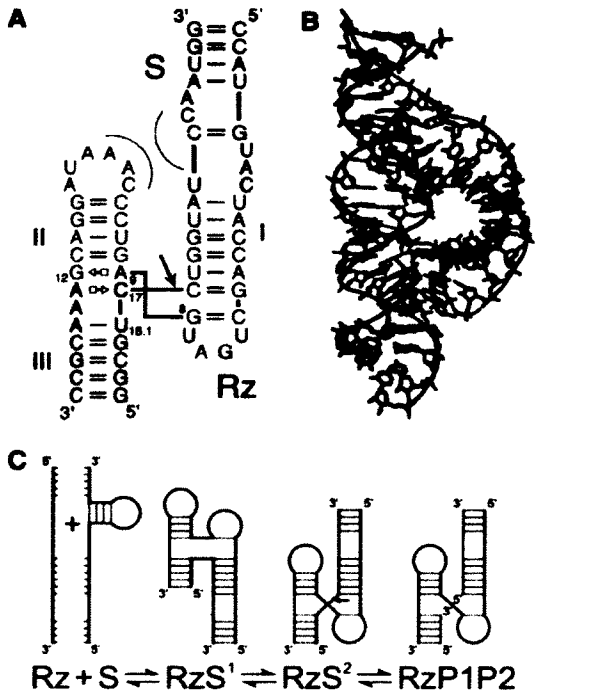


Fig. 3 Structure and folding pathway of the hammerhead ribozyme. The ribozyme and substrate are denoted by *Rz* and *S*, respectively. The cleavage site is indicated by an *arrow*. **a** Secondary structure of a *trans*-acting version of the hammerhead ribozyme [72] according to the notation of Leontis and Westhof [145]. The corresponding minimal version is obtained by removing the shaded region at the top. The tertiary interactions between the loop of stem II and the bulge of stem I are illustrated by the *thin curved lines*. **b** Crystal structure of the *Schistosome* version of hammerhead ribozyme [72]. The *colors* of the various domains are the same in panels A and B. **c** Hammerhead ribozyme's folding pathway. The substrate is shown in *green*, and the products in *red*

Hammerhead ribozyme

The hammerhead ribozyme is undoubtedly the most studied catalytic RNA so far. It was the first to be discovered [18, 66] and to have its crystal structure determined [67, 68]. It is famous for the dilemma surrounding the poor correlation that exists between the biochemical data and the crystal structures of minimal versions [69–71]. The hammerhead ribozyme adopts a "Y-shape" consisting of three helical stems of variable sequence and length named stems I–III (Fig. 3a). In the center of the three-way junction of the Y, 11 highly conserved nucleotides form the catalytic core that includes the cleavage site located 3′ of an important single-stranded cytosine residue [17]. A recent X-ray crystal structure of a larger version derived from *Schistosoma mansoni* underlines the importance of two peripheral elements, specifically a hairpin loop located at the end of stem II and a bulge located in the longer stem I (Fig. 3b) [72]. Interaction of these two motifs in a tertiary manner increases the cleavage rate by at least 50-fold [73]. Most, if not all, natural hammerhead self-cleaving motifs appear to contain similar interactions between loops or bulges located in stems I and II [74, 75]. Moreover, this new crystal structure reconciled older chemical data and provided a basis to hypothesize how the transition state might be achieved [76]. In this transition state an interaction network implicating key nucleotides (G8, A9, G12, U16.1 and C17) positions the 2′-hydroxyl nucleophile, the scissile phosphate and the 5′-oxygen leaving group at an angle very close to 180° [72].

Trans-acting hammerhead ribozymes have been extensively used to develop gene-inactivation systems either for functional genomic applications or for therapeutic aims [77, 78]. Several *trans*-acting versions have been produced by cleaving the RNA strand at various locations within the

loops, with the most common involving strand breaks in loops I and III. In such *trans*-acting versions, the initial step of the folding pathway is the formation of the ribozyme–substrate complex via assembly of both stems I and III. Then the tertiary interaction network is established, which enables formation of the transition state for cleavage (Fig. 3c). Depending on the length and the base composition of the stems formed between the substrate and the ribozyme, product release inhibition has been observed.

The hammerhead ribozyme is also by far the most studied RNA molecule in terms of the modulation of the catalytic activity, either directly or by the addition of motifs leading to the production of allosteric versions. Several compounds were reported to either positively or negatively modulate the hammerhead cleavage activity [79–81]. For example, in the case of antibiotics, usually inhibitory effects were observed. This is nicely exemplified by the effect of neomycin on hammerhead activity [79]. In this case, it has been proposed that neomycin interacts directly with the ribozyme–substrate complex, stabilizing the ground state over the transition state and thereby inhibiting cleavage. Other antibiotics, such as chlorotetracycline, either stimulate or inhibit depending on the reaction conditions [80]. However, in general the mechanisms by which antibiotics affect the cleavage activity remain elusive [79]. Another interesting mechanism has been reported in a study of metal ion inhibition of the hammerhead cleavage. For example, terbium ions tightly bind to the ribozyme and subsequently displace an adjacent magnesium cation important for the catalytic activity [82]. In the case of cobalt hexammine, a study revealed that it binds to the hammerhead ribozyme and provokes a conformational change that produces an inactive structure [83]. Finally, several proteins and peptides with RNA chaperone activities have been shown to positively modulate hammerhead ribozyme activity [84–89]. For example, the p7 nucleocapsid protein of HIV has been demonstrated to resolve misfolded ribozyme–substrate complexes [84]. Other proteins, such as hnRNP A1, enhance the binding rate of the substrate and significantly accelerate product release, thereby increasing ribozyme turnover [85].

Since Watson–Crick base pairs are predictable, several strategies directed towards controlling hammerhead activity were based on the use of either additional sequences or complementary oligonucleotides. For example, in the case of modulation through additional sequences, the SOFA module developed for the HDV ribozyme was adapted to the hammerhead ribozyme [62] (Fig. 4a). The blocker was inserted at the 5′ end of stem I and interacts with the 3′ arm of stem III, thereby blocking the ribozyme in an "off" state. The biosensor complementary to the 5′ region of the substrate was added to the 3′ end of stem III. Substrate annealing to the biosensor then promotes the displacement

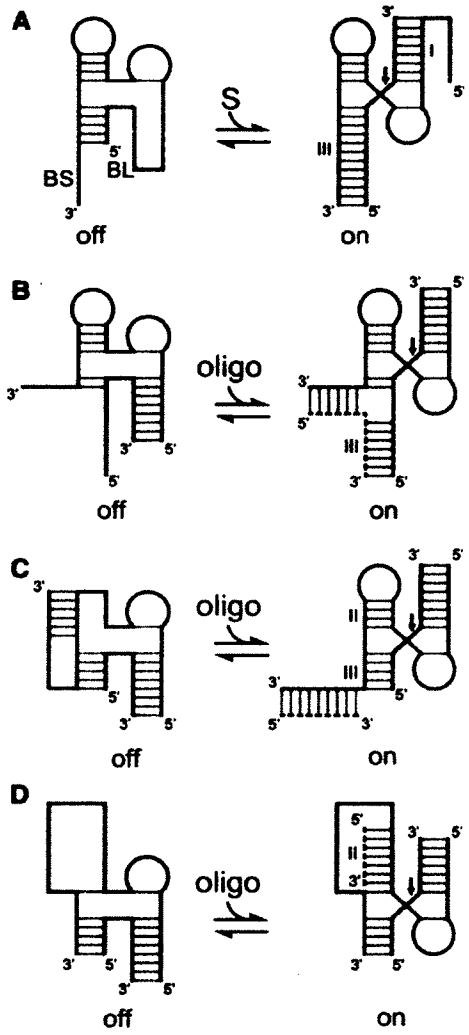


Fig. 4 Various modulation strategies for the hammerhead ribozyme based on complementary sequences. The inactive and active states are indicated by "*off*" and "*on*", respectively. The substrate is shown in *green* and the portion of the ribozyme responsible for the modulation in *blue*. The *trans*-acting oligonucleotides are represented by the *dotted lines*. **a** The SOFA module [62]. *BL* and *BS* indicate blocker and biosensor, respectively. **b** Three-strand ribozyme strategy [91]. **c** Blocker release strategy [93]. **d** Stem II stabilization strategy [92]

of the blocker, yielding a ribozyme in an "on" state. In the case of modulation through complementary oligonucleotides, facilitator oligonucleotides that interact with regions of the substrate outside of the binding site of the ribozyme were shown to modify the ribozyme's kinetic parameters. Positive effects are seen when the facilitator binds the substrate 3′ of the ribozyme, while negative ones are observed when it binds 5′ of the ribozyme [90]. Alternatively, other strategies involving *trans*-acting oligonucleotides that interact with the ribozyme were designed in order to modulate the cleavage activity (Fig. 4b–d) [91–94].

Table 1 Allosteric hammerhead ribozymes

Effector	Modulation	Region	Mechanism	Ref.
Allosteric hammerhead ribozymes developed by rational design				
ATP	–	Stem II	Steric interference between bound aptamer and HHRz tertiary structure	[96]
FMN	+	Stem II	Stabilization of stem II	[97]
ATP and FMN	+	Stem III	Control the accessibility/formation of the substrate binding site	[99]
ATP and theophylline	±	Stem II	Stabilization or disruption of stem II	[95]
FMN and theophylline	+	Stem II	Stabilization of stem II	[98]
FMN and theophylline	+	Stem II	Stabilization of stem II	[100]
Theophylline, tetracycline and xanthine	±	Stem II	In vivo RNA-based gene-regulatory platform	[106]
TMPyP4	±	Stem II	Stabilization of stem II	[101]
Light and caged theophylline	+	Stem II	Stabilization of stem II	[105]
ERK2 phosphorylated or unphosphorylated	+	Stem II	Stabilization of stem II	[102]
HCV helicase/replicase	+	Stems I, II, III and I,III	Increase the affinity of the HHRz for its substrate	[103]
Tat protein	±	Stem II	Stabilization of stem II or slip-structure mechanism	[104]
Allosteric hammerhead ribozymes developed by in vitro selection				
Oligonucleotides	+	Stem II	Reorganization and stabilization of stem II	[114]
cAMP, cCMP and cGMP	+	Stem II	Slip-structure mechanism	[107] [108]
FMN, theophylline and ATP	±	Stem II	Slip-structure mechanism	[113]
Theophylline and 3-methylxanthine	+	Stem II	The level of modulation depend on the communication module selected	[110]
Caffeine and aspartame	+	Stem II	n.d.	[109]
Mn^{2+}, Fe^{2+}, Co^{2+}, Ni^{2+}, Zn^{2+} and Cd^{2+}	+	Stem II	Stabilization of stem II	[112]
Light and (BDHP-COOH or BCPD-COOH)	+	Stem II	Stabilization of stem II	[111]

n.d. Not determined

Over the years, rational designs and in vitro selection strategies have been used for the development of a wide repertoire of allosteric hammerhead ribozymes (also named aptazymes). The ribozyme activity is switched from inactive to active, or vice versa, by the binding of an effector to an adjacent aptamer domain (Table 1, upper portion) [95–106]. The general strategy involves using aptamers previously reported to be specific for a compound and then inserting them in one of the structural stems of the hammerhead ribozyme. Insertion in either stem I or III generally regulates the substrate binding step, while insertion in stem II usually either impairs the formation or decreases the stability of this helical region critical for the catalytic activity. Aptamers can be specific to various small components and proteins, including adenosine-5′-triphosphate (ATP), theophylline, flavin mononucleotide (FMN) and the HIV Tat protein, to name only a few examples, and have been reported to regulate hammerhead catalytic activity [95–100, 104]. Interestingly, two allosteric ribozymes using the insertion of the same ATP aptamer within stem II have been demonstrated to regulate the catalysis in different ways [95, 96]. The only difference between the two was in the linker domain located between the aptamer domain and the catalytic core. In one construct this communication module is composed of 4 base pairs that form a proper stem II in the absence of ATP, leading to negative modulation [95, 96]. In the second construct, the presence of a linker composed of 14 bases including only 3 base pairs prevents the formation of stem II and yields an inactive ribozyme in the absence of ATP, leading to positive modulation [95]. Clearly, this illustrates the importance of the communication module in the allosteric behavior of a ribozyme. Moreover, cooperativity between the binding of two different effectors has also been reported [100]. In this case, FMN and theophylline aptamers have been juxtaposed within stem II and sequential binding of both effectors is required in order to activate the ribozyme.

Several allosteric hammerhead ribozymes have been isolated by in vitro selection (see Table 1, lower portion) [107–114]. Generally, in terms of allosteric ribozymes, the initial rounds are performed using a negative selection process that removes the sequences that do not require the presence of the effector for cleavage to occur, followed by subsequent rounds using a positive selection in the presence of the effector. For most of the reported allosteric hammerhead ribozymes, the region including the randomized positions was introduced in stem II. In vitro selection yielded allosteric domains that exhibit a very high level of specificity for their effectors, even discriminating between closely related compounds such as cAMP, cCMP and cGMP [107, 108]. Alternatively, the reselection of an allosteric hammerhead ribozyme with an improved discrimination power for a given effector after mutagenesis of the original aptamer domain has also been reported [110]. This type of strategy has been used to isolate ribozymes whose activities are modulated by photosensitive compounds [105, 111], opening the way to in vivo light-regulated gene expression through the activity of photochemically modulated hammerhead ribozymes.

Finally, as is the case with the rational design of allosteric ribozymes, the domain linking the aptamer and the catalytic core is very important, in terms of the modulation, for in vitro selected allosteric ribozymes. In vitro selection of a communication module connected to an FMN-aptamer has shown that in some cases the effector causes a positive modulation, while for other sequence variants it causes a negative regulation [113]. In a similar study, several theophylline-dependent allosteric hammerhead ribozymes harboring various sequences as communication sequences have been isolated, illustrating the flexibility of this domain [110]. Furthermore, a communication module is not restricted to one specific aptamer or particular ribozyme. The same activating communication module selected for the FMN aptamer was used in a construction with both an ATP aptamer and a theophylline aptamer, causing a positive modulation of the activity in the presence of the ligand. In the same way, a communication module selected with the HDV ribozyme was also effective when transferred into a hammerhead ribozyme [65]. Clearly, the communication module is a key feature of allostreric ribozymes, permitting the transfer of properties from the aptamer domain to the catalytic core. The molecular mechanism leading to the modulation of in vitro selected hammerhead ribozymes remains largely unknown, mainly because the efforts to elucidate the structures of both the inactive and active conformation were limited. In many cases, the stabilization of stem II has been proposed to be required for the adoption of a catalytically active structure.

Hairpin ribozyme

In its naturally occurring form, the hairpin ribozyme adopts a four-way junction secondary structure [115, 116]. Both two-way and four-way junction *trans*-acting ribozymes have been derived and extensively studied (Fig. 5a) [117–119]. In the minimal two-way junction version, the 49-nucleotide-long ribozyme is composed of four helices separated into two domains (A and B). Domain A comprises stems I and II (or H1 and H2) and an internal loop of eight nucleotides, while domain B comprises stems III and IV (or H3 and H4) and an internal loop of 16 nucleotides. These two domains dock into a non-coaxial orientation, and the interactions between the two internal loops form the catalytic core of the ribozyme (Fig. 5b) [120]. This catalytic core can be formed with a variety of domain-connecting setups; in fact, it can even be formed if the two domains are separated into two distinct RNA molecules [121, 122]. The cleavage site and almost all of the conserved nucleotides are located in these two internal loops

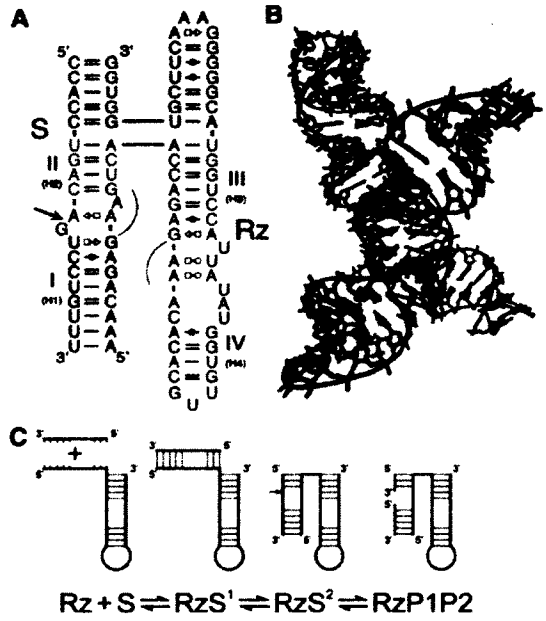


Fig. 5 Structure and folding pathway of the hairpin ribozyme. The ribozyme and substrate are denoted by *Rz* and *S*, respectively. The cleavage site is indicated by an *arrow*. **a** Secondary structure of a four-way junction *trans*-acting version of the hairpin ribozyme [116] according to the notation of Leontis and Westhof [145]. The corresponding minimal two-way version is obtained by removing the shaded region at the top. The tertiary interactions between the internal loops A and B are illustrated by the *thin curved lines*. The harmonized nomenclature using roman numbers for the stems is indicated as well as the previous one in *parenthesis*. **b** Crystal structure of the four-way junction version [116]). The *colors* of the various domains are the same in panels A and B. **c** Hairpin ribozyme's folding pathway. The substrate is shown in *green* and the products in *red*

[123]. The presence of a guanosine 3′ of the cleavage site is absolutely required as it is involved in stacking tertiary interactions within the internal loop of domain B, being a key feature of the hairpin catalytic core [124]. In contrast, a variety of mutations can be found in the base-paired regions as long as they support helix formation, the only requirement being the presence of a guanosine as the first nucleotide in the ribozyme strand of stem II [125].

Folding of the hairpin ribozyme is a two-step folding pathway: after the binding of the substrate that results in the formation of domain A, the docking of the two internal loops caused by tertiary interactions forms the catalytic site (Fig. 5c). Once it is formed, the chemical step can then occur using in-line geometry. Interestingly, the ligation reaction is around tenfold more efficient than the cleavage reaction [126]. The hypothesis to explain this observation is that the docking of the two internal loops confers sufficient rigidity to the catalytic core that the optimal geometry is maintained and the highly reactive 2′–3′-cyclic phosphate can open and react with the 5′-hydroxyl [126].

At least four different ways of modulating the activity of the hairpin ribozyme have been succesfully used. First, in vitro selected adenine-dependent hairpin ribozymes have been developed; those require the metabolite adenine as a structural component to be able to form their catalytic cores (Fig. 6a) [127]. Second, the replacement of the short stem IV with a communicator module has permitted the introduction of a redox active aptamer into the ribozyme resulting in allosteric regulation of the ribozyme's activity (Fig. 6b) [128]. Third, a SOFA version was developed in which a short blocker sequence forming a stem II equivalent was attached to the 3′end of the ribozyme, and a biosensor recognition sequence that is able to bind the substrate was added to the 5′end (Fig. 6c) [62]. As with the hammerhead ribozyme, the binding of the substrate to the biosensor allows the substrate to unfold the blocker and bind in its place, thereby increasing the specificity of the reaction. Finally, a modulation based on the presence of a *trans*-acting oligonucleotide has been achieved by introducing a short stem loop between stems II and III. When an oligonucleotide composed of either DNA or RNA complementary to this new stem loop is bound, domains A and B can no longer interact together and the catalytic activity is disrupted (Fig. 6d) [129, 130].

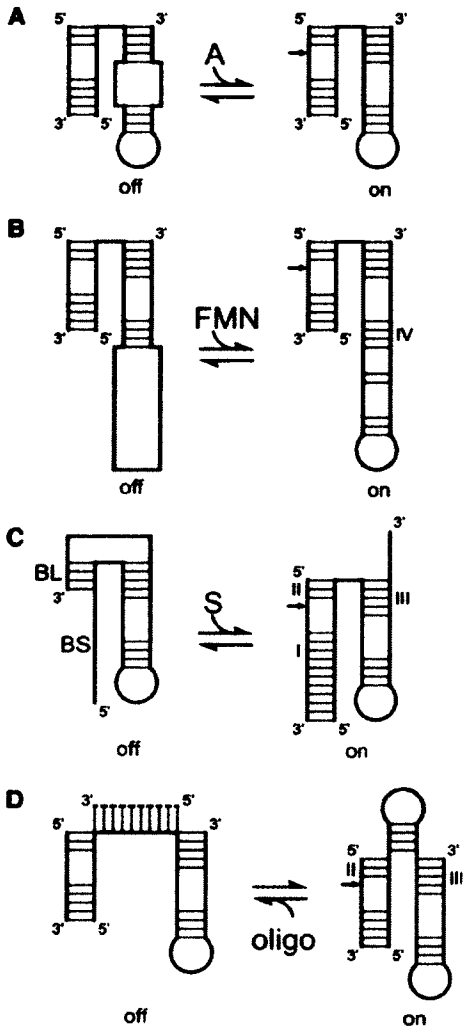


Fig. 6 Various modulations of the activity of the hairpin ribozyme. The inactive and active states are indicated by "*off*" and "*on*", respectively. The substrate is shown in *green*, and the portion of the ribozyme responsible for the modulation in *blue*. **a** The adenosine-dependant hairpin ribozyme [127]. **b** The FMN aptamer attached by a communication module [128]. **c** The SOFA module [62]. *BL* and *BS* indicate blocker and biosensor, respectively. **d** The oligonucleotide modulation [129]. The *trans*-acting oligonucleotide is represented by the *dotted line*

VS ribozyme

With a length of around 140 nucleotides in its *trans*-acting version, the VS ribozyme is the largest known small self-cleaving ribozyme. Its crystal structure has not yet been solved, although a recent small-angle X-ray scattering study gave a low-resolution structure, which was used to fit the ribozyme and to confirm its predicted overall shape [131]. The secondary structure of this ribozyme is formed by 5 helices, numbered from II to VI, that form two three-way junctions sharing stem III (Fig. 7a). The substrate is composed of stem I and, in contrast to other ribozymes, the VS ribozyme is able to rely exclusively on tertiary interactions in order to recognize its substrate [132]. The substrate appears to be docked between stems II and VI, in a cleft formed by the three-way junction involving stems II, III and VI, where it makes close interactions with the internal loop of stem VI (Fig. 7b). An adenosine (A756) in this highly conserved internal loop and a guanosine (G638)

in the internal loop of the substrate have been shown to be important for catalysis and are probably directly implicated in the chemical reaction, leading to the hypothesis that these internal loops are in fact the ribozyme's active site [133]. The cleavage occurs in the internal loop of stem I through a general acid–base catalyzed transesterification mechanism. This cleavage has been shown to be a consequence of a one nucleotide shift in the stable secondary structure of stem I [134], a secondary structure rearrangement that is stabilized by the formation of a kissing loop pseudoknot interaction. This pseudoknot is composed of three base pairs formed between the loops of stems I and V, and is required for the cleavage by VS ribozyme [135]. Hydroxyl radical footprinting showed that a solvent-inaccessible core is rapidly formed in this region [136]; the absence of water is supported by a previous kinetic study that reported pH variation has only a small effect on this catalytic RNA [137]. The folding of the VS ribozyme is ion-induced, and the presence of divalent ions is required, while the presence of monovalent ions further enhances the reaction rate [137].

While screening a collection of antibiotics looking for inhibitors of the VS ribozyme, a class of simple cyclic peptides was found to enhance ribozyme cleavage [138]. The presence of tuberactinomycin antibiotics and, especially, viomycin, in a *trans*-cleavage experiment results in a more extensive reaction and a decrease in the magnesium requirements for VS ribozyme cleavage. These antibiotics are able to interact with RNA and, in the case of the VS ribozyme, to restore the activity of some mutants, although the exact mechanism is not known.

Finally, a longer version of the natural VS ribozyme has a non-essential stem (called stem VII) that increases the stabilization between the 3′ end of the ribozyme and the 5′ end of its substrate (Fig. 7a). This version is particularly interesting in ligation assays since it permits substrate binding through a secondary structure [139].

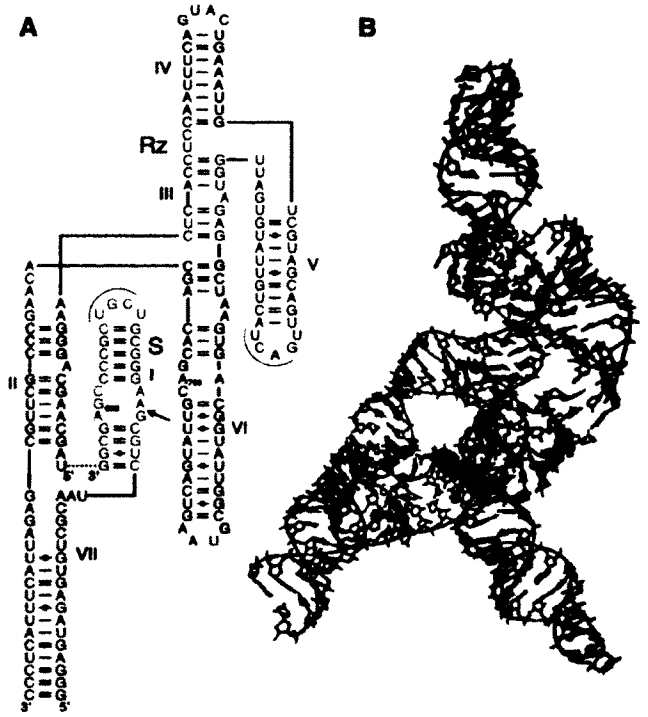


Fig. 7 Structure of the VS ribozyme. The ribozyme and substrate are denoted by *Rz* and *S*, respectively. The cleavage site is indicated by an *arrow*. **a** Secondary structure of a *trans*-acting version of the VS ribozyme according to the notation of Leontis and Westhof [145]. A shortened version can be obtained by removing the non-essential shaded region at the bottom. The tertiary interactions forming the kissing loop between loops I and V are illustrated by the *thin curved lines*. The *dotted line* represents the junction that has been removed in order to generate a *trans*-acting version. **b** Structure of the VS ribozyme obtained by SAXS [131]. The *colors* of the various domains are the same in panels A and B

GlmS ribozyme

The *trans*-acting version of the *glmS* ribozyme is around 120 nucleotides long and is composed of eight small stems, two of which are pseudoknots (Fig. 8a). The available crystal structures have shown that coaxial stacking of RNA helices defines the global tertiary structure of the *glmS* ribozyme, while the catalytic core is formed by an internal loop involved in the two pseudoknots (Fig. 8b) [140, 141]. The double pseudoknot catalytic core forms a rigid binding pocket that is able to recognize and use a variety of small hydroxylamines, such as ethanolamine or tris-hydroxymehylaminomethane, although only binding of glucosamine results in efficient cleavage [142]. The *glmS* ribozyme does not appear to undergo important conformational changes upon the binding of its cofactor glucosamine-6-phosphate (GlcN6P). When the cofactor GlcN6P is bound in the catalytic site through the recognition of the sugar moiety and of the phosphate [143], 80% of the solvent-accessible surface of the metabolite is buried, and the primary amine is well positioned for catalysis, either directly as a general acid, or as a general base through water molecules [140]. Metal ions are required for the formation of the precleavage structure, but do not appear to play any catalytic roles in *glmS* ribozyme activity [143].

Even if *glmS* ribozyme activity does not rely on conformational changes, its unique metabolite dependence implies an extremely interesting and as yet poorly understood folding pathway. The binding site for the GlcN6P metabolite is formed through a complex network of interactions that provides sufficient rigidity to the pocket so that it stays open, with the functional groups ready to

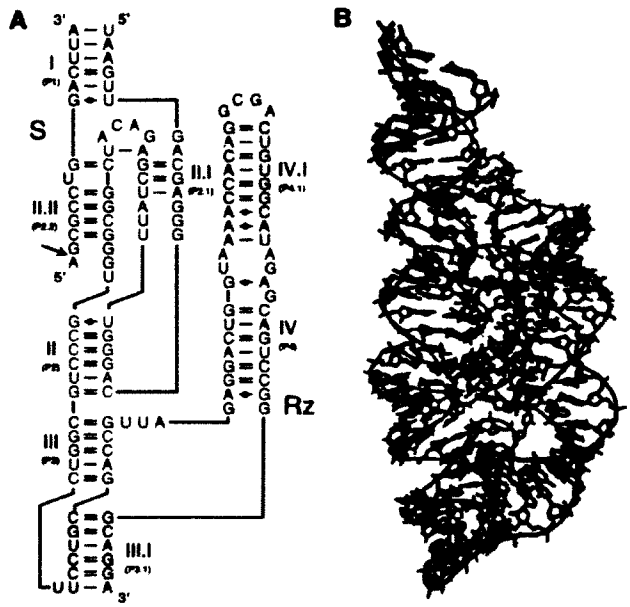


Fig. 8 Structure of the *glmS* ribozyme. The ribozyme and substrate are denoted by *Rz* and *S*, respectively. The cleavage site is indicated by an arrow. **a** Secondary structure of a *trans*-acting version of the *glmS* ribozyme according to the notation of Leontis and Westhof [145]. The harmonized nomenclature using roman numbers for the stems is indicated as well as the previous one in *parentheses*. **b** Crystal structure of the *glmS* ribozyme from *Thermoanaerobacter tengcongensis* [140]. The *colors* of the various domains are the same in panels A and B

accommodate the cofactor, quite a different scenario from the induced fit folding usually seen upon substrate binding with other RNA enzymes. Most of the research to date has focused on understanding the cleavage reaction, but very little is known about the folding steps leading to the cofactor-free structure. This ribozyme has only recently been discovered, and already some progress has been made toward the elucidation of its folding pathway. In addition, both the structural and functional versatility of this ribozyme has been examined by in vitro selection [144].

Concluding remarks

Studies on the natural small ribozymes have made key contributions to our understanding of the folding of RNA molecules, and have helped to identify structural features important for the formation of catalytically active structures. Although the accurate prediction of the secondary and tertiary structures of long RNA species is not yet possible, the current knowledge obtained from these small ribozymes has considerably improved our view of the folding pathways of RNA molecules in general in addition to providing opportunities for the development of a substantial number of methodologies for their study. Moreover, the discovery of other small, natural, self-catalytic ribozymes and riboswitches will widen the repertoire of this class of catalytic ribozyme and will most likely increase our ability to correctly predict their folding. The characterization of the interactions involved, combined with the various components modulating ribozyme activity, has paved the way for many developments in the area of the nanotechnology. Progress in comprehending how the various RNA motifs interplay together in order to control activity should lead to a better understanding of complex biological mechanisms such as the regulation of translation involving several RNA motifs that form the 5′ untranslated region of mRNA. From this point of view, RNA molecules can be considered as combinations of building blocks. The designing of complex ribozymes with several motifs modulating either together, or sequentially, the catalytic activity is expected to propel further progress in this direction.

Acknowledgments The work in J.P.P.'s laboratory is supported by a grant from the Canadian Institute for Health Research (CIHR: MOP-44022). The RNA group is supported by grants from the CIHR and the Université de Sherbrooke. J.P.P. holds the Canada Research Chair in Genomics and Catalytic RNA and is a member of the Infectious Diseases group of the Centre de Recherche Clinique Étienne-Lebel.

References

1. Serganov A, Patel DJ (2007) Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. Nat Rev Genet 8:776–790
2. Onoa B, Tinoco I Jr (2004) RNA folding and unfolding. Curr Opin Struct Biol 14:374–379
3. Zhuang Z, Jaeger L, Shea JE (2007) Probing the structural hierarchy and energy landscape of an RNA T-loop harpin. Nucleic Acids Res 35:6995–7002
4. Greenleaf WJ, Frieda KL, Foster DA, Woodside MT, Block SM (2008) Direct observation of hierarchical folding in single riboswitch aptamers. Science 319:630–633
5. Moody EM, Bevilacqua PC (2003) Folding of a stable DNA motif involves a highly cooperative network of interactions. J Am Chem Soc 125:16285–16293
6. Jaeger L, Verzemnieks EJ, Geary C (2009) The UA-handle: a versatile submotif in stable RNA architectures. Nucleic Acids Res 37:215–230
7. Draper DE (2008) RNA folding: thermodynamic and molecular descriptions of the roles of ions. Biophys J 95:5489–5495
8. Reymond C, Bisaillon M, Perreault JP (2009) Monitoring of an RNA multistep folding pathway by isothermal titration calorimetry. Biophys J 96:132–140
9. Fiore J, Kraemer B, Koberling F, Erdmann R, Nesbitt D (2009) Enthalpy-driven RNA folding: single-molecule thermodynamics of tetraloop receptor tertiary interaction. Biochemistry 48:2550–2558

10. Pljevaljcic G, Klostermeier D, Millar DP (2005) The tertiary structure of the hairpin ribozyme is formed through a slow conformational search. Biochemistry 44:4870–4876
11. Russell R, Das R, Suh H, Travers KJ, Laederach A, Engelhardt MA, Herschlag D (2006) The paradoxical behavior of a highly structured misfolded intermediate in RNA folding. J Mol Biol 363:531–544
12. Bhaskaran H, Russell R (2007) Kinetic redistribution of native and misfolded RNAs by a DEAD-box chaperone. Nature 449:999–1000
13. Alemán EA, Lamichhane R, Rueda D (2008) Exploring RNA folding one molecule at a time. Curr Opin Chem Biol 12:647–654
14. Lilley DM (2005) Structure, folding and mechanisms of ribozymes. Curr Opin Struct Biol 15:313–323
15. Cochrane JC, Strobel SA (2008) Catalytic strategies of self-cleaving ribozymes. Acc Chem Res 41:1027–1035
16. Symons RH (1997) Plant pathogenic RNAs and RNA catalysis. Nucleic Acids Res 25:2683–2689
17. Hutchins CJ, Tathjen PD, Forster AC, Symons RH (1986) Self-cleavage of plus and minus transcripts of avocado sunblotch viroid. Nucleic Acids Res 14:3627–3640
18. Prody GA, Bakos JT, Buzayan JM, Schneider IR, Bruening G (1986) Autolytic processing of dimeric plant virus satellite RNA. Science 231:1577–1580
19. Buzayan JM, Gerlach WL, Bruening G (1986) Non-enzymatic cleavage and ligation of RNAs complementary to a plant virus satellite RNA. Nature 323:349–353
20. Earnshaw DJ, Gait MJ (1997) Progress toward the structure and therapeutic use of the hairpin ribozyme. Antisense Nucleic Acid Drug Dev 7:403–411
21. Lilley DM (1999) Folding and catalysis by the hairpin ribozyme. FEBS Lett 452:26–30
22. Buzayan JM, Gerlach WL, Bruening G (1986) Nonenzymatic cleavage and ligation of RNAs complementary to a plant virus satellite RNA. Nature 323:349–353
23. Shih IH, Been MD (2002) Catalytic strategies of the hepatitis delta virus ribozymes. Annu Rev Biochem 71:887–917
24. Been MD (1994) Cis- and trans-acting ribozymes from a human pathogen, hepatitis delta virus. Trends Biochem Sci 19:251–256
25. Been MD, Wickham GS (1997) Self-cleaving ribozymes of hepatitis delta virus RNA. Eur J Biochem 247:741–753
26. Epstein LM, Gall JG (1987) Self-cleaving transcripts of satellite DNA from the newt. Cell 48:535–543
27. Ferbeyre G, Smith JM, Cedergren R (1998) Schistosome satellite DNA encodes active hammerhead ribozymes. Mol Cell Biol 18:3880–3888
28. Rojas AA, Vazquez-Tello A, Ferbeyre G, Venanzetti F, Bachmann L, Paquin B, Sbordoni V, Cedergren R (2000) Hammerhead-mediated processing of satellite pDo500 family transcripts from Dolichopoda cave crickets. Nucleic Acids Res 28:4037–4043
29. Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW (2006) A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. Science 313:1788–1792
30. Saville BJ, Collins RA (1990) A site-specific self-cleavage reaction performed by a novel RNA in Neurospora mitochondria. Cell 61:685–696
31. Collins RA (2002) The Neurospora Varkud satellite ribozyme. Biochem Soc Trans 30:1122–1126
32. Barrick JE, Corbino KA, Winkler WC, Nahvi A, Mandal M, Collins J, Lee M, Roth A, Sudarsan N, Jona I et al (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. Proc Natl Acad Sci USA 101:6421–6426
33. Winkler WC, Nahvi A, Roth A, Collins JA, Breaker RR (2004) Control of gene expression by a natural metabolite-responsive ribozyme. Nature 428:263–264
34. Cochrane JC, Strobel SA (2008) Catalytic strategies of self-cleaving ribozymes. Acc Chem Res 41:102701035
35. Murray JB, Seyhan AA, Walter NG, Burke JM, Scott WG (1998) The hammerhead, hairpin and VS ribozymes are catalytically proficient in monovalent cations alone. Chem Biol 5:587–595
36. Perrotta AT, Been MD (2006) HDV ribozyme activity in monovalent cations. Biochemistry 45:11357–11365
37. Deschênes P, Ouellet J, Perreault J, Perreault JP (2003) Formation of the P1.1 pseudoknot is critical for both the cleavage activity and substrate specificity of an antigenomic trans-acting hepatitis delta ribozyme. Nucleic Acids Res 31:2087–2096
38. Nishikawa F, Shirai M, Nishikawa S (2002) Site-specific modification of functional groups in genomic hepatitis delta virus (HDV) ribozyme. Eur J Biochem 269:5792–5803
39. Nehdi A, Perreault JP (2006) Unbiased in vitro selection reveals the unique character of the self-cleaving antigenomic HDV RNA sequence. Nucleic Acids Res 34:584–592
40. Ferré-D'Amaré AR, Zhou K, Doudna JA (1998) Crystal structure of a hepatitis delta virus ribozyme. Nature 395:567–574
41. Wadkins TS, Perrotta AT, Ferré-D'Amaré AR, Doudna JA, Been MD (1999) A nested double pseudoknot is required for self-cleavage activity of both the genomic and antigenomic hepatitis delta virus ribozymes. RNA 5:720–727
42. Suh YA, Kumar PK, Taira K, Nishikawa S (1993) Self-cleavage activity of the genomic HDV ribozyme in the presence of various divalent metal ions. Nucleic Acids Res 21:3277–3280
43. Ferré-D'Amaré AR, Doudna JA (2000) Crystallization and structure determination of a hepatitis delta virus ribozyme: use of the RNA-binding protein U1A as a crystallization module. J Mol Biol 295:541–556
44. Tanaka Y, Hori T, Tagaya M, Katahira M, Nishikawa F, Sakamoto T, Kurihara Y, Nishikawa S, Uesugi S (2000) NMR analysis of tertiary interactions in HDV ribozymes. Nucleic Acids Symp Ser 44:285–286
45. Fiola K, Perreault JP (2002) Kinetic and binding analysis of the catalytic involvement of ribose moieties of a trans-acting delta ribozyme. J Biol Chem 277:26508–26516
46. Ke A, Zhou K, Ding F, Cate JH, Doudna JA (2004) A conformational switch controls hepatitis delta virus ribozyme catalysis. Nature 429:201–205
47. Perrotta AT, Been MD (1990) The self-cleaving domain from the genomic RNA of hepatitis delta virus: sequence requirements and the effects of denaturant. Nucleic Acids Res 18:6821–6827
48. Duhamel J, Liu DM, Evilia C, Fleysh N, Dinter-Gottlieb G, Lu P (1996) Secondary structure content of the HDV ribozyme in 95% formamide. Nucleic Acids Res 24:3911–3917
49. Reymond C, Ouellet J, Bisaillon M, Perreault JP (2007) Examination of the folding pathway of the antigenomic hepatitis delta virus ribozyme reveals key interactions of the L3 loop. RNA 13:44–54
50. Ouellet J, Perreault JP (2004) Cross-linking experiments reveal the presence of novel structural features between a hepatitis delta virus ribozyme and its substrate. RNA 10:1059–1072
51. Ananvoranich S, Lafontaine DA, Perreault JP (1999) Mutational analysis of the antigenomic trans-acting delta ribozyme: the alterations of the middle nucleotides located on the P1 stem. Nucleic Acids Res 27:1473–1479
52. Nishikawa F, Nishikawa S (2000) Requirement for canonical base pairing in the short pseudoknot structure of genomic hepatitis delta virus ribozyme. Nucleic Acids Res 28:925–931
53. Nehdi A, Perreault J, Beaudoin JD, Perreault JP (2007) A novel structural rearrangement of hepatitis delta virus antigenomic ribozyme. Nucleic Acids Res 35:6820–6831
54. Harris DA, Rueda D, Walter NG (2002) Local conformational changes in the catalytic core of the trans-acting hepatitis delta

virus ribozyme accompany catalysis. Biochemistry 41:12051–12061
55. Perrotta AT, Wadkins TS, Been MD (2006) Chemical rescue, multiple ionizable groups, and general acid–base catalysis in the HDV genomic ribozyme. RNA 12:1282–1291
56. Nakano S, Chadalavada DM, Bevilacqua PC (2000) General acid–base catalysis in the mechanism of a hepatitis delta virus ribozyme. Science 287:1493–1497
57. Peracchi A (2004) Prospects for antiviral ribozymes and deoxyribozymes. Rev Med Virol 14:47–64
58. Bergeron LJ, Perreault JP (2005) Target-dependent on/off switch increases ribozyme fidelity. Nucleic Acids Res 33:1240–1248
59. Bergeron LJ, Reymond C, Perreault JP (2005) Functional characterization of the SOFA delta ribozyme. RNA 11:1858–1868
60. Robichaud GA, Perreault JP, Ouellette RJ (2008) Development of an isoform-specific gene suppression system: the study of the human Pax-5B transcriptional element. Nucleic Acids Res 36:4609–4620
61. Fiola K, Perreault JP, Cousineau B (2006) Gene targeting in the Gram-Positive bacterium Lactococcus lactis, using various delta ribozymes. Appl Env Microbiol 72:869–879
62. Lévesque D, Brière FP, Perreault JP (2007) A modern mode of activation for nucleic acid enzymes. PLoS ONE 2:e673. doi: 10.1371/journal.pone.0000673
63. Lévesque D, Beaudoin JD, Roy S, Perreault JP (2007) In vitro selection and characterization of RNA aptamers binding thyroxine hormone. Biochem J 403:129–138
64. Beaudoin JD, Perreault JP (2008) Potassium ions modulate a G-quadruplex-ribozyme's activity. RNA 14:1018–1025
65. Kertsburg A, Soukup GA (2002) A versatile communication module for controlling RNA folding and catalysis. Nucleic Acids Res 30:4599–4606
66. Forster AC, Symons RH (1987) Self-cleavage of plus and minus RNAs of a virusoid and a structural model for the active sites. Cell 49:211–220
67. Pley HW, Flaherty KM, McKay DB (1994) Three-dimensional structure of a hammerhead ribozyme. Nature 372:68–74
68. Scott WG, Finch JT, Klug A (1995) The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. Cell 81:991–1002
69. McKay DB (1996) Structure and function of the hammerhead ribozyme: an unfinished story. RNA 2:395–403
70. Verma S, Vaish NK, Eckstein F (1997) Structure–function studies of the hammerhead ribozyme. Curr Opin Chem Biol 1:532–536
71. Blount KF, Uhlenbeck OC (2005) The structure–function dilemma of the hammerhead ribozyme. Annu Rev Biophys Biomol Struct 34:415–440
72. Martick M, Scott WG (2006) Tertiary contacts distant from the active site prime a ribozyme for catalysis. Cell 126:309–320
73. Canny MD, Jucker FM, Kellogg E, Khvorova A, Jayasena SD, Pardi A (2004) Fast cleavage kinetics of a natural hammerhead ribozyme. J Am Chem Soc 126:10848–10849
74. De la Peña M, Gago S, Flores R (2003) Peripheral regions of natural hammerhead ribozymes greatly increase their self-cleavage activity. EMBO J 22:5561–5570
75. Khvorova A, Lescoute A, Westhof E, Jayasena SD (2003) Sequence elements outside the hammerhead ribozyme catalytic core enable intracellular activity. Nat Struct Biol 10:708–712
76. Nelson JA, Uhlenbeck OC (2008) Hammerhead redux: does the new structure fit the old biochemical data? RNA 14:605–615
77. Amarzguioui M, Prudz H (1998) Hammerhead ribozyme design and application. Cell Mol Life Sci 54:1175–1202
78. Citti L, Rainaldi G (2005) Synthetic hammerhead ribozymes as therapeutic tools to control disease genes. Curr Gene Ther 5:11–24
79. Stage TK, Hertel KJ, Uhlenbeck OC (1995) Inhibition of the hammerhead ribozyme by neomycin. RNA 1:95–101
80. Murray JB, Arnold JR (1996) Antibiotics interactions with the hammerhead ribozyme: tetracyclines as a new class of hammerhead inhibitor. Biochem J 317:860–885
81. Jenne A, Hartig JS, Piganeau N, Tauer A, Samarsky DA, Green MR, Davies J, Fumalok M (2001) Rapid identification and characterization of hammerhead-ribozyme inhibitors using fluorescence-based technology. Nat Biotechnol 19:56–61
82. Feig AL, Scott WG, Uhlenbeck OC (1998) Inhibition of the hammerhead ribozyme cleavage reaction by site-specific binding of Tb(III). Science 279:81–84
83. Horton TE, DeRose VJ (2000) Cobalt hexammine inhibition of the hammerhead ribozyme. Biochemistry 39:11408–11416
84. Hearschlag D, Khosla M, Tsuchihashi Z, Karpel RL (1994) An RNA chaperone activity of non-specific RNA binding proteins in hammerhead ribozyme catalysis. EMBO J 13:2913–2924
85. Bertrand EL, Rossi JJ (1994) Facilitation of hammerhead ribozyme catalysis by the nucleocapsid protein of HIV-1 and the heterogeneous nuclear ribonucleoprotein A1. EMBO J 13:2904–2912
86. Sioud M, Jespersen L (1996) Enhancement of hammerhead ribozyme catalysis by glyceraldehyde-3-phosphate dehydrogenase. J Mol Biol 257:775–789
87. Tsuchihashi Z, Khosla M, Herschlag D (1993) Protein enhancement of hammerhead ribozyme catalysis. Science 262:99–102
88. Kuciak M, Gabus C, Ivanyi-Nagy R, Semrad K, Storchak R, Chaloin O, Muller S, Mély Y, Darlix JL (2008) The HIV-1 transcriptional activator Tat has potent nucleic acid chaperoning activities in vitro. Nucleic Acids Res 36:3389–3400
89. Huang ZS, Wu HN (1998) Identification and characterization of the RNA chaperone activity of hepatitis delta antigen peptides. J Biol Chem 273:26455–26461
90. Jankowsky E, Schwenger B (1996) Oligonucleotide facilitators may inhibit or activate a hammerhead ribozyme. Nucleic Acids Res 24:423–429
91. Wang DY, Lai BH, Feldman AR, Sen D (2002) A general approach for the use of oligonucleotide effectors to regulate the catalysis of RNA-cleaving ribozymes and DNAzymes. Nucleic Acids Res 30:1735–1742
92. Porta H, Lizardi PM (1995) An allosteric hammerhead ribozyme. Biotechnology 13:161–164
93. Burke DH, Ozerova ND, Nilsen-Hamilton M (2002) Allosteric hammerhead ribozyme TRAPs. Biochemistry 41:6588–6594
94. Komatsu Y, Yamashita S, Kazama N, Nobuoka K, Ohtsuka E (2000) Construction of new ribozymes requiring short regulator oligonucleotides as a cofactor. J Mol Biol 229:1231–1243
95. Tang J, Breaker RR (1997) Rational design of allosteric ribozymes. Chem Biol 4:453–459
96. Tang J, Breaker RR (1998) Mechanism for allosteric inhibition of an ATP-sensitive ribozyme. Nucleic Acids Res 26:4214–4221
97. Araki M, Okuno Y, Hara Y, Sugiura Y (1998) Allosteric regulation of a ribozyme activity through ligand-induced conformational change. Nucleic Acids Res 26:3379–3384
98. Soukup GA, Breaker RR (1999) Design of allosteric hammerhead ribozymes activated by ligand-induced structure stabilization. Structure 15:783–791
99. Wang DY, Lai BH, Sen D (2002) A general strategy for effector-mediated control of RNA-cleaving ribozymes and DNA enzymes. J Mol Biol 318:33–43
100. Jose AM, Soukup GA, Breaker RR (2001) Cooperative binding of effectors by an allosteric ribozyme. Nucleic Acids Res 29:1631–1637

101. Wieland M, Hartig JS (2006) Turning inhibitors into activators: a hammerhead ribozyme controlled by a guanine quadruplex. Angew Chem Int Ed Engl 45:5875–5878
102. Vaish NK, Dong F, Andrews L, Schweppe RE, Ahn NG, Blatt L, Seiwert SD (2002) Monitoring post-translational modification of proteins with allosteric ribozymes. Nat Biotechnol 20:810–815
103. Cho S, Kim JE, Lee BR, Kim JH, Kim BG (2005) Bis-aptazyme sensors for hepatitis C virus replicase and helicase without blank signal. Nucleic Acids Res 33:e177
104. Wang DY, Sen D (2002) Rationally designed allosteric variants of hammerhead ribozymes responsive to the HIV-1 Tat protein. Comb Chem High Throughput Screen 5:301–312
105. Young DD, Deiters A (2006) Photochemical hammerhead ribozyme activation. Bioorg Med Chem Lett 16:2658–2661
106. Win MN, Smolke CD (2007) A modular and extensible RNA-based gene-regulatory platform for engineering cellular function. Proc Natl Acad Sci USA 104:14283–14288
107. Koizumi M, Soukup GA, Kerr JN, Breaker RR (1999) Allosteric selection of ribozymes that respond to the second messengers cGMP and cAMP. Nat Struct Biol 6:1062–1071
108. Marshall KA, Ellington AD (1999) Training ribozymes to switch. Nat Struct Biol 6:992–994
109. Ferguson A, Boomer RM, Kurz M, Keene SC, Diener JL, Keefe AD, Wilson C, Cload ST (2004) A novel strategy for selection of allosteric ribozymes yields RiboReporterTM sensors for caffeine and aspartame. Nucleic Acids Res 32:1756–1766
110. Soukup GA, Emilsson GA, Breaker RR (2000) Altering molecular recognition of RNA aptamers by allosteric selection. J Mol Biol 298:623–632
111. Lee HW, Robinson SG, Bandyopadhyay S, Mitchell RH, Sen D (2007) Reversible photo-regulation of a hammerhead ribozyme using a diffusible effector. J Mol Biol 371:1163–1173
112. Zivarts M, Liu Y, Breaker RR (2005) Engineered allosteric ribozymes that respond to specific divalent metal ions. Nucleic Acids Res 33:622–631
113. Soukup GA, Breaker RR (1999) Engineering precision RNA molecular switches. Proc Natl Acad Sci USA 96:3584–3589
114. Komatsu Y, Nobuoka K, Karino-Abe N, Matsuda A, Ohtsuka E (2002) In vitro selection of hairpin ribozymes activated with short oligonucleotides. Biochemistry 41:9090–9098
115. Hampel A, Tritz R (1989) RNA catalytic properties of the minimum (-)sTRSV sequence. Biochemistry 28:4929–4933
116. Rupert PB, Ferré-D'Amaré AR (2001) Crystal structure of a hairpin ribozyme–inhibitor complex with implications for catalysis. Nature 410:780–786
117. Fedor MJ (2000) Structure and function of the hairpin ribozyme. J Mol Biol 297:269–291
118. Ferré-D'Amaré AR (2004) The hairpin ribozyme. Biopolymers 73:71–78
119. Wilson TJ, Nahas M, Araki L, Harusawa S, Ha T, Lilley DM (2007) RNA folding and the origins of catalytic activity in the hairpin ribozyme. Blood Cells Mol Dis 38:8–14
120. Komatsu Y, Koizumi M, Nakamura H, Ohtsuka E (1994) Loop-size variation to probe a bent structure of a hairpin ribozyme. J Am Chem Soc 116:3692–3696
121. Komatsu Y, Kanzaki I, Ohtsuka E (1996) Enhanced folding of hairpin ribozymes with replaced domains. Biochemistry 35:9815–9820
122. Butcher SE, Heckman JE, Burke JM (1995) Reconstitution of hairpin ribozyme activity following separation of functional domains. J Biol Chem 270:29648–29651
123. Berzal-Herranz A, Joseph S, Burke JM (1992) In vitro selection of active hairpin ribozymes by sequential RNA-catalyzed cleavage and ligation reactions. Genes Dev 6:129–134
124. Chowrira BM, Berzal-Herranz A, Burke JM (1991) Novel guanosine requirement for catalysis by the hairpin ribozyme. Nature 354:320–322
125. Hampel A, Tritz R, Hicks M, Cruz P (1990) "Hairpin" catalytic RNA model: evidence for helices and sequence requirement for substrate RNA. Nucleic Acids Res 18:299–304
126. Fedor MJ (1999) Tertiary structure stabilization promotes hairpin ribozymes ligation. Biochemistry 38:11040–11050
127. Meli M, Vergne J, Maurel MC (2003) In vitro selection of adenine-dependent hairpin ribozymes. J Biol Chem 278:9835–9842
128. Strohbach D, Novak N, Müller S (2006) Redox-active riboswitching: allosteric regulation of ribozyme activity by ligand-shape control. Angew Chem Int Ed Engl 45:2127–2129
129. Najafi-Shoushtari SH, Famulok M (2007) DNA aptamer-mediated regulation of the hairpin ribozyme by human alpha-thrombin. Blood Cells Mol Dis 38:19–24
130. Vauléon S, Müller S (2003) External regulation of hairpin ribozyme activity by an oligonucleotide effector. Chembiochem 4:220–224
131. Lipfert J, Ouellet J, Norman DG, Doniach S, Lilley DM (2008) The complete VS ribozyme in solution studied by small-angle X-ray scattering. Structure 16:1357–1367
132. Guo HC, Collins RA (1995) Efficient trans-cleavage of a stem-loop RNA substrate by a ribozyme derived from neurospora VS RNA. EMBO J 14:368–376
133. Lilley DM (2004) The Varkud satellite ribozyme. RNA 10:151–158
134. Rastogi T, Beattie TL, Olive JE, Collins RA (2006) A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. EMBO J 15:2820–2825
135. Andersen AA, Collins RA (2000) Rearrangement of a stable RNA secondary structure during VS ribozyme catalysis. Mol Cell 5:469–478
136. Hiley SL, Collins RA (2001) Rapid formation of a solvent-inaccessible core in the Neurospora Varkud satellite ribozyme. EMBO J 20:5461–5469
137. Collins RA, Olive JE (1993) Reaction conditions and kinetics of self-cleavage of a ribozyme derived from Neurospora VS RNA. Biochemistry 32:2795–2799
138. Olive JE, De Abreu DM, Rastogi T, Andersen AA, Mittermaier AK, Beattie TL, Collins RA (1995) Enhancement of Neurospora VS ribozyme cleavage by tuberactinomycin antibiotics. EMBO J 14:3247–3251
139. Jones FD, Ryder SP, Strobel SA (2001) An efficient ligation reaction promoted by a Varkud Satellite ribozyme with extended 5′- and 3′-termini. Nucleic Acids Res 29:5115–5120
140. Klein DJ, Ferré-D'Amaré AR (2006) Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. Science 313:1745–1747
141. Cochrane JC, Lipchock SV, Strobel SA (2007) Structural investigation of the GlmS ribozyme bound to its catalytic cofactor. Chem Biol 14:97–105
142. McCarthy TJ, Plog MA, Floy SA, Jansen JA, Soukup JK, Soukup GA (2005) Ligand requirements for glmS ribozyme self-cleavage. Chem Biol 12:1221–1226
143. Roth A, Nahvi A, Lee M, Jona I, Breaker RR (2006) Characteristics of the glmS ribozyme suggest only structural roles for divalent metal ions. RNA 12:607–619
144. Link KH, Guo L, Breaker RR (2006) Examination of the structural and functional versatility of glmS ribozymes by using in vitro selection. Nucleic Acids Res 34:4968–4975
145. Leontis NB, Westhof E (2001) Geometric nomenclature and classification of RNA base pairs. RNA 7:499–512

Nucleic Acids Research Advance Access published October 11, 2007

Nucleic Acids Research, 2007, 1–12
doi:10.1093/nar/gkm674

A novel structural rearrangement of hepatitis delta virus antigenomic ribozyme

Atef Nehdi, Jonathan Perreault, Jean-Denis Beaudoin and Jean-Pierre Perreault*

RNA Group/Groupe ARN, Département de Biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, J1H 5N4, Canada

Received May 7, 2007; Revised August 17, 2007; Accepted August 18, 2007

ABSTRACT

A bioinformatic covariation analysis of a collection of 119 novel variants of the antigenomic, self-cleaving hepatitis delta virus (HDV) RNA motif supported the formation of all of the Watson–Crick base pairs (bp) of the catalytic centre except the C19–G81 pair located at the bottom of the P2 stem. In fact, a novel Watson–Crick bp between C19 and G80 is suggested by the data. Both chemical and enzymatic probing demonstrated that initially the C19–G81 pair is formed in the ribozyme (Rz), but upon substrate (S) binding and the formation of the P1.1 pseudoknot C19 switches its base-pairing partner from G81 to G80. As a result of this finding, the secondary structure of this ribozyme has been redrawn. The formation of the C19–G80 bp results in a J4/2 junction composed of four nucleotides, similar to that seen in the genomic counterpart, thereby increasing the similarities between these two catalytic RNAs. Additional mutagenesis, cleavage activity and probing experiments yield an original characterization of the structural features involving the residues of the J4/2 junction.

INTRODUCTION

Both the genomic and antigenomic hepatitis delta virus (HDV) RNA strands include a self-cleaving RNA motif that produces 2′–3′-cyclic phosphate and 5′-hydroxyl termini (1,2). These self-cleaving RNAs have been separated into two molecules in order to develop *trans*-acting systems in which one molecule, the ribozyme (Rz), possesses the catalytic properties required to successively cleave several molecules of substrate (S). The HDV ribozyme folds into a double-pseudoknot secondary structure composed of one stem (P1), two pseudoknots (P1.1 and P2), two stem-loops (P3–L3 and P4–L4) and three single-stranded junctions (J1/2, J1/4 and J4/2) (Figure 1). Crystallographic studies of the HDV genomic ribozyme have provided high-resolution details of the compact tertiary structure of this catalytic RNA (3,4).

The J1/4 junction and the L3 loop are both initially single-stranded, but subsequently are involved in the formation of the P1.1 pseudoknot (5–7). The formation of this pseudoknot, which requires the presence of both the substrate and magnesium, is critical for the folding of the ribozyme into an active ternary structure (8). It permits the stacking of the P1–P1.1–P4 stem into one helix that becomes coaxial to the second stacked P2–P3 helix (3,4). Moreover, it significantly contributes to the bringing together of the scissile phosphate and the catalytic cytosine that is located within the J4/2 junction (i.e. C_{76}) (9,10). The nucleotides of this single-stranded region are also involved in two distinct structural motifs: a ribose zipper formed between the two adenosines of the J4/2 junction (A78 and A79) and the two cytosines of the P3 stem (C21 and C22), and a trefoil turn structure formed by the catalytic cytosine (C76) and the adjacent guanosine (G_{77}) (3). The formation of this trefoil turn positions C_{76} deep within the catalytic core near the scissile bond (3,11).

In a recent study, the primary results of an unbiased *in vitro* selection of antigenomic HDV ribozymes randomized at 25 positions within the catalytic centre (Figure 1) showed that nucleotide variation was found at all of the randomized positions, even those where the specific base in question was believed to be essential for catalytic activity (12). Analysis of the random nucleotide covariation, obtained using a database composed of 45 different sequences, supported the formation of most of ribozyme base pairs that form both the P1 and the P3 stems. Moreover, these results support the presence of the homopurine bp located at the top of the P4 stem. Altogether, these observations led us to conclude that the selection performed was unbiased and yielded many new variants. However, neither new base pairs, nor any tertiary interactions, were discovered. This might be due to the fact that the number of different variants was relatively small (only 45).

As a result, we decided to emphasize both the sequencing and the analysis of the sequence variations.

*To whom correspondence should be addressed. Tel: +1 819 564 5310; Fax: +1 819 564 5340; Email: jean-pierre.perreault@usherbrooke.ca

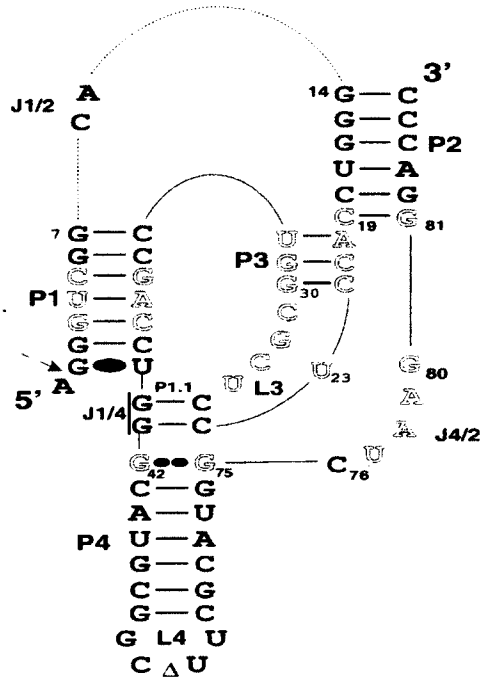


Figure 1. Secondary structure of the antigenomic self-cleaving HDV RNA. The numbering system used is from Shih and Been (34). The empty letters indicate the 25 position randomized in the previous study (12). The Wobble bp is represented by a single large dot (G•U), while the homopurine bp at the top of the P4 stem is represented by two large dots (G••G). The triangle in the L4 loop indicates the P4 deletion (as compared to the natural variants). The arrow indicates the cleavage site. The *trans*-acting ribozyme was created by deleting the CA forming the J1/2 junction.

This work permitted the identification of a new structural rearrangement involving the J4/2 junction and the bottom base pair of the of the P2 stem. This conformational transition increases the similarities of the geometry adopted by the J4/2 junctions of both the genomic and antigenomic HDV ribozymes.

MATERIALS AND METHODS

Sequence determination and covariation analysis

A total of 170 new, self-cleaving HDV motif clones, isolated from the previously reported *in vitro* selection (12), were sequenced using the T7 sequencing kit (GE Healthcare). All of the sequences from both the present and the previous studies were compiled in a database and subsequently analysed for base conservation both at each position and in terms of nucleotide covariation using homemade software written in Perl language (Figure 2). Briefly, the software analyses two positions at a time, namely X and Y. First, for a given position X, the percentages of the different nucleotides found at position Y in all sequences are determined ($\%Y_AX_N$,$\%Y_C$ X_N,$\%Y_GX_N$ and$\%Y_UX_N$). Second, for a specific nucleotide in position X (for example a cytosine), the percentage of each nucleotide

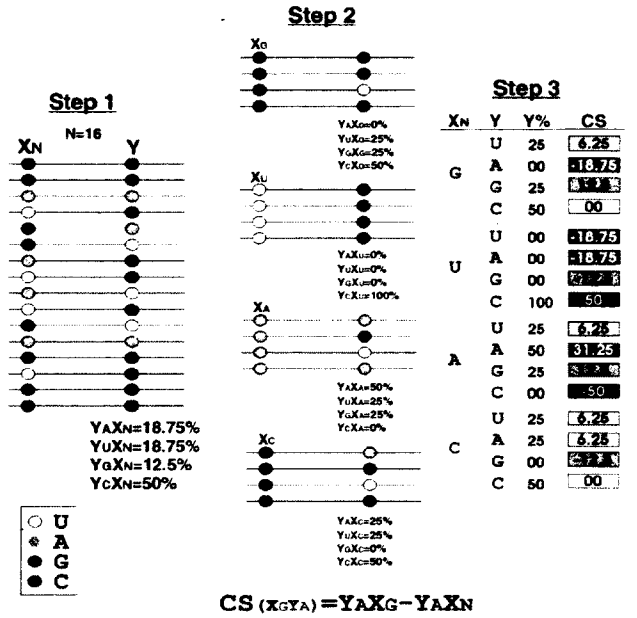


Figure 2. Schematic representation of the covariation software developed for the analysis of the sequence database. The three steps of the procedure are illustrated using the example described in the text. CS indicates the covariation score. The inset shows the legend of the filling sphere for each nucleotide.

(G, C, A and U) at position Y is determined ($\%Y_AX_C$,$\%Y_CX_C$,$\%Y_GX_C$,$\%Y_UX_C$). Third, these two percentages are compared in order to determine a relative covariation score (CS) for the specific nucleotides located in these two positions [$CS = (\%Y_N' - \%Y_N)$: $CS(Y_AX_C) = \%Y_AX_C - \%\ Y_AX_C$]. If the CS is 0, or near 0, this suggests that the nucleotide in position X has no effect on the distribution of nucleotides in position Y. Conversely, a relatively large CS (e.g. ±0.5) suggests a significant covariation of these two bases (see Supplementary data for examples). The operation was repeated for all possible combinations of the 25 randomized positions (i.e. 9600 combinations), and the CSs were displayed in an Excel file.

RNA synthesis and ^{32}P-5′-end labelling

Chemical synthesis. The substrates, the SdA4 analogue, the 3′ product and the RzB strands (with and without the abasic residue at position 77) were chemically synthesized using 2′-ACE chemistry by Dharmacon Research Inc. (Lafayette, Colorado). The resulting polymers were deprotected according to the manufacturer's recommendations and then purified by denaturing 10–20% polyacrylamide gels electrophoresis (PAGE, 19:1 ratio of acrylamide to bisacrylamide) using 45 mM Tris–borate, pH 7.5, 8 M urea and 1 mM EDTA solution as buffer. The products were visualized by UV shadowing, the bands corresponding to the correct sizes were cut out and the RNA eluted by incubating overnight at room temperature in a solution containing 0.1% SDS and 0.5 M ammonium acetate. The resulting polymers were ethanol precipitated

and their quantities determined by absorbance at 260 nm after dissolving in water.

In vitro *transcription*. All of the ribozymes and the RzA strands were synthesized by run-off transcription as described previously (19). Briefly, RNA molecules were produced by annealing two strands of complementary DNA oligonucleotides that included the T7 RNA promoter followed by the sequence of the desired ribozyme. The transcriptions were then performed in the presence of 10 µg purified T7 RNA polymerase, 24 U RNA Guard (Amersham Biosciences), 0.01 U pyrophosphatase (Roche Diagnostics) and 500 pmol of template in a buffer containing 80 mM HEPES–KOH pH 7.5, 24 mM $MgCl_2$, 2 mM spermidine, 40 mM DTT and 5 mM of each NTP in a final volume of 100 µl at 37°C for 2 h. Upon completion, the reaction mixtures were treated with DNase RQ1 (Amersham Biosciences) at 37°C for 20 min. The RNA was then purified by phenol extraction and ethanol precipitation, and was fractionated by denaturing 8% PAGE. The bands corresponding to the correct sizes for the RNA species were cut out, the transcripts eluted and then ethanol precipitated.

^{32}P-5′-end labelling. Purified substrates or ribozymes (40 pmol) were dephosphorylated in a final volume of 10 µl using 10 U of Antartic phosphatase according to the manufacturer's conditions (New England Biolabs). The reactions were stopped by heating for 5 min at 65°C. Dephosphorylated RNAs (10 pmol) were 5′-end labelled in a final volume of 10 µl containing 3.2 pmol [γ-^{32}P]-ATP (6000 Ci/mmol, Amersham Biosciences), 10 mM Tris–HCl, pH 7.5, 10 mM $MgCl_2$, 50 mM KCl and 3 U of T4 polynucleotide kinase (Amersham Biosciences) at 37°C for 90 min. The reactions were stopped by the addition of formamide dye buffer and the mixtures fractionated through denaturing 10–20% PAGE gels. After autoradiography, the bands containing the appropriate 5′-end-labelled RNAs were excised and the RNA recovered as described above.

Chemical and enzymatic probing

Ribozymes were subjected to in-line probing according to a procedure reported previously (16). Briefly, 1 nM of ^{32}P-5′-end-labelled *trans*-acting ribozyme was incubated for 40 h at 25°C in a buffer containing 20 mM $MgCl_2$, 50 mM Tris–HCl pH 8.3 and 100 mM KCl either in the absence or the presence of SdA4 analogue (20 µM). In the enzymatic digestions, trace amounts of the 5′-end-labelled ribozymes (<1 nM) were dissolved in 10 µl of buffer containing 20 mM Tris–HCl, pH 7.5, 10 mM $MgCl_2$ and 100 mM NH_4Cl. The mixtures were incubated for 0.5–1.0 min at 25°C in the presence of either 0.2 U of RNase T1 (Amersham Biosciences) or of 0.001 U of RNase V1 (Pierce Molecular Biology), and were then quenched by adding 10 µl of formamide. For alkaline hydrolysis, the ribozymes (<1 nM) were dissolved in 4 µl of water and 2 µl of 2 N NaOH were added. The reaction was incubated at room temperature for 15 s, and was then quenched by the addition of 6 µl of 500 mM Tris–HCl, pH 7.5 and 5 µl of loading buffer. The resulting mixtures were ethanol precipitated, resuspended in loading buffer, separated on denaturating 10% PAGE gels and visualized by exposure of the gels to phosphor imaging screens.

Cleavage reactions and kinetic assays. The cleavage reactions were performed as described previously (19). Briefly, cleavage reactions were carried out in 20 µl reaction mixtures containing 50 mM Tris–HCl, pH 7.5 and 10 mM $MgCl_2$ at 37°C under single-turnover conditions ($[Rz] \gg [S]$). Prior to the reaction, trace amounts of 5′-end-labelled substrate (<1 nM) and non-radioactive ribozymes (100 nM) were mixed together, heated at 70°C for 1 min, snap-cooled on ice for 3 min and then incubated at 37°C for 5 min. Similarly in the case of the bimolecular constructs (RzA–RzB or–RzB–Ab77) 100 nM of each RNA species were used in each reaction. Following this pre-incubation step, the cleavage reactions were initiated by the addition of $MgCl_2$. Aliquots (2 µl) were removed either at various times up to 1 h, or until the end point of the cleavage was reached, and were quenched by the addition of ice-cold formamide dye buffer (8 µl). The mixtures were fractionated on denaturing 20% PAGE gels and exposed to phosphor Imager screens (Molecular Dynamics). The extent of cleavage was determined from measurements of the radioactivity present both in the substrate and in the 5′ product bands at each time point using the ImageQuant software. When required, the rate of cleavage (k_{obs}) was obtained by fitting the data to the equation $A_t = A_\infty\ (1 - e^{-kt})$, where A_t is the percentage of cleavage at time t, A_∞ is the maximum percent cleavage (or the end point of cleavage) and k is the rate constant (k_{obs}). Each rate constant was calculated from at least two independent measurements.

RESULTS

Analysis of all selected HDV variants

The different experiments reported in the previous study were performed using a database containing a total of 330 clones, but involving only 45 different sequence variants (12). In order to find novel variants, we decided to sequence 170 additional clones that were obtained via a PCR amplification strategy that avoided domination by the most active self-cleaving ribozymes. A new database containing 119 different variants was constructed (Supplementary data Table 1). Subsequently, each variant was individually synthesized by run-off transcription, and its self-cleavage activity determined (Supplementary data Table 1). All of the selected variants were active, confirming the accuracy of the selection protocol. Among the 119 variants, 21 sequences exhibited self-cleavage activities at a level smaller than 10% after 60 min of incubation. The 98 most active variants were retained for further analysis in order to study the interactions important for efficient self-cleavage.

In order to perform a covariation analysis of the randomized nucleotides, we developed a software program that analyses the covariation between the randomized positions. This program can detect not only Watson–Crick bp, but also any tertiary interactions (Figure 2).

Table 1. Compilation of data from the covariation analysis

Positions		Combinations of nucleotides (%)																Base-pairing (%)	
		Mismatchs										Waston-Crick bp				G-U:U-G Wobble		Waston-Crick bp	G/U:U/G Wobble
		CC	GG	UU	AA	AC	AG	GA	UC	CU	CA	AU	UA	GC	CG	UG	GU		
P1 stem	(3–37)	2	11	2	0	1	4	0	1	5	2	4	0	17	18	3	30	39	33
	(4–36)	2	15	2	0	0	2	11	0	3	3	1	0	9	29	3	20	39	23
	(5–35)	2	2	1	0	3	1	0	3	0	0	3	2	29	35	1	18	69	19
P3 stem	(20–32)	0	0	0	0	0	1	0	0	0	1	80	10	5	0	0	3	95	3
	(21–31)	0	0	0	0	0	0	0	0	0	1	1	11	5	82	0	0	99	0
	(22–30)	0	1	0	0	0	0	0	0	0	0	3	18	21	56	1	0	98	1
Homopurine	(42–75)	0	17	1	79	0	2	1	0	0	0	0	0	0	0	0	0	0	0
Bottom of the P2 stem	**(19–80)**	1	0	1	0	0	1	1	0	1	0	42	13	21	17	1	1	**93**	**2**
	(19–81)	0	18	0	15	3	24	3	13	3	2	0	0	1	14	3	1	**15**	**4**

Briefly, a CS is calculated for each pair of nucleotides. A CS that tends near zero indicates there is no covariation between these two nucleotides. A significant positive CS indicates that the presence of a cytosine in position X favours the presence of an adenosine in position Y, and vice versa. Conversely, a significant negative difference is an indication that covariation is disfavoured.

Table 1 presents a compilation of the CSs of random nucleotides known to interact together based on the secondary structure of the antigenomic HDV ribozyme. The covariation results are in good agreement with most of the interactions proposed in the secondary structure of the HDV ribozyme (1). For example, the base pairs of the P3 stem are well supported by the covariation results: each base pair is found to be formed in at least 98% of the selected ribozymes, with a significant preference for Watson–Crick base pairing over GU/UG Wobble bp (≥95% compared to ≤3%). This is in good agreement with the results of earlier directed mutagenesis studies which indicated that the size of the P3 stem is critical to self-cleaving RNA strands, and that the identities of its nucleotides are also important (13). Similarly, covariation analysis supported the formation of the base pairs located in the middle of the P1 stem (see Table 1 and Supplementary data Table 2). However, statistical analysis showed that these base pairs (G_3C_{37}, U_4A_{36} and C_5G_{35}) tolerate the presence of more Wobble bp and mismatches than do those of the P3 stem (Table 1). This indicates that they are important, but not essential, at least for self-cleaving RNA strands. The fact that mismatches are observed in the middle of the P1 stem is in agreement with the results from enzymatic probing experiments that illustrated the susceptibility of this domain to a single-strand specific ribonuclease (RNase T2) (14). The presence of the homopurine bp found at the top of the P4 stem (positions 42 and 75) is well supported by the covariation analysis. AA and GG homopurines were retrieved in 96% of the sequences, while AG and GA were found in only 3%. Surprisingly, the selected sequences predominantly include an AA homopurine bp (79%). This contrasts with the exclusive presence of a GG homopurine bp in the sequence variants found in nature (15). This difference might be due to different selective pressures existing *in vivo* as compared to *in vitro*. Thus, the covariation analysis confirmed the secondary structure of the self-cleaving antigenomic sequences, including both the homopurine and all Watson–Crick bp, with one exception (see below).

The base pairing of the nucleotides at the bottom of the P2 stem (C_{19}–G_{81}) was believed to form a Watson–Crick bp (Figure 1). The covariation analysis does not support the presence of this base pair (see Table 1 and Supplementary data Table 2). According to the set of selected sequences, only 19% of the selected variants possess either a Watson–Crick or Wobble bp at this position (Table 1). CSs attributed to the different nucleotides that can form Watson–Crick bp are either very close to zero, or negative (e.g. $A_{19} - U_{81} = 0$; $U_{19} - A_{81} = -0.2$; $G_{19} - C_{81} = -0.1$; $C_{19} - G_{81} = 0.08$). In order to add biochemical support to this observation, *trans*-acting mutant ribozymes, including the three possible mutations at position 81, were synthesized and their cleavage activities assessed under single-turnover conditions. The rate constants (k_{obs}) of the mutants were determined, and are presented in histogram form (Figure 3). The least active mutant was the wild-type ribozyme harbouring a C_{19}–G_{81} bp. The three mutants that did not permit the formation of a base pair were all more active. Independent experiments with several other mutants at the bottom of the P2 stem (positions 19 and 81) also led to the same conclusion. Even several ribozymes without a base pair at this position were found to be more active than the wild-type ribozyme (Ouellet,J. and Perreault,J.P, unpublished data).

The analysis of the nucleotide occupying positions 19 and 80 gave high CSs suggesting that these nucleotides were able to form a Watson–Crick bp (e.g. $A_{19} - U_{80} = 0.43$; $U_{19} - A_{80} = 0.57$; $G_{19} - C_{80} = 0.69$; $C_{19} - G_{80} = 0.63$) (see Supplementary data Table 2). In fact, 93% of the selected sequences possess nucleotides at these positions that have the ability to form a Watson–Crick bp, and 2% a Wobble bp, between these two residues (Table 1). In other words, 95% of the sequence variants seemed to include a base pair at these positions. Conversely, only 15% of the selected ribozymes were able to form a base pair between positions 19 and 81.

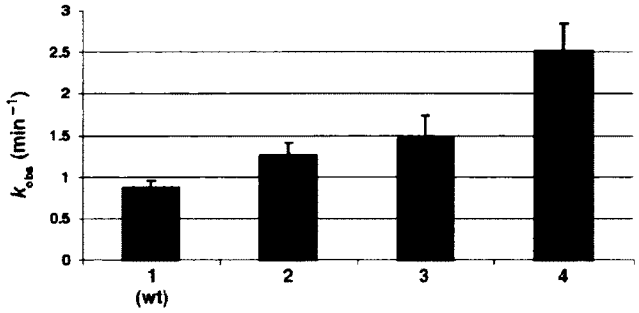


Figure 3. Histogram of the determined rate constant (k_{obs}) for the wild-type ribozyme (1), the $RzG_{81}A$ (2), the $RzG_{81}C$ (3) and the $RzG_{41}U$ and (4) mutants.

All possible base-pairing combinations were found, but at different levels (A_{19}–U_{80}, U_{19}–A_{80}, G_{19}–C_{80} and C_{19}–G_{80} were found at 42%, 13%, 21% and 17%, respectively). To our knowledge, this is the first demonstration of a base pairing between these two specific nucleotides.

Probing of the bottom of the P2 stem

In-line probing is a method that relies on the fact that there is a natural rate of spontaneous cleavage within RNAs (16). Cleavage occurs when a phosphodiester linkage is subjected to an internal nucleophilic attack by the 2′ oxygen adjacent to and in-line with it. Structured regions of RNA, such as those in the base-paired stems, are less susceptible to spontaneous cleavage than are non-structured regions (17). In-line probing experiments were performed with the *trans*-acting version of the HDV ribozyme for which the kinetic behaviour has been extensively characterized under both single- and multiple-turnover conditions (18). The use of a *trans*-acting version permits the monitoring of the binding of the substrate to the ribozyme, a step which has been shown to be essential for many conformational transitions to take place (5,19,20). Probing experiments were performed using a trace amount (<1 nM) of ^{32}P 5′-end-labelled ribozyme in either the presence, or the absence, of an excess of uncleavable substrate (SdA4; [S] ≫ [Rz]). The use of a substrate analogue that includes a deoxyriboadenine adjacent to the cleavage site prevents the cleavage from occurring. In the absence of SdA4, RNA degradation was observed in all single-stranded regions, including the nucleotides of both the P1 region (positions 33–39) and the J1/4 junction (G_{40} and G_{41}) that are not base paired under these conditions (Figure 4A, lane 1; see also Supplementary Figure S1A). The nucleotides of the J4/2 junction, including G_{80}, appear to be single-stranded. Because the residue G_{81} does not appear to be hydrolysed, this indicates that it is base paired with C_{19}. Upon the addition of SdA4, several modifications in the banding pattern were observed (Figure 4A, lane 2; see also Supplementary Figure 1A). In general, the intensities of the majority of the bands were weaker, indicating that the presence of the substrate led to a more compact structure. Specifically, the nucleotides of the P1, P1.1 and P4 stems appeared to

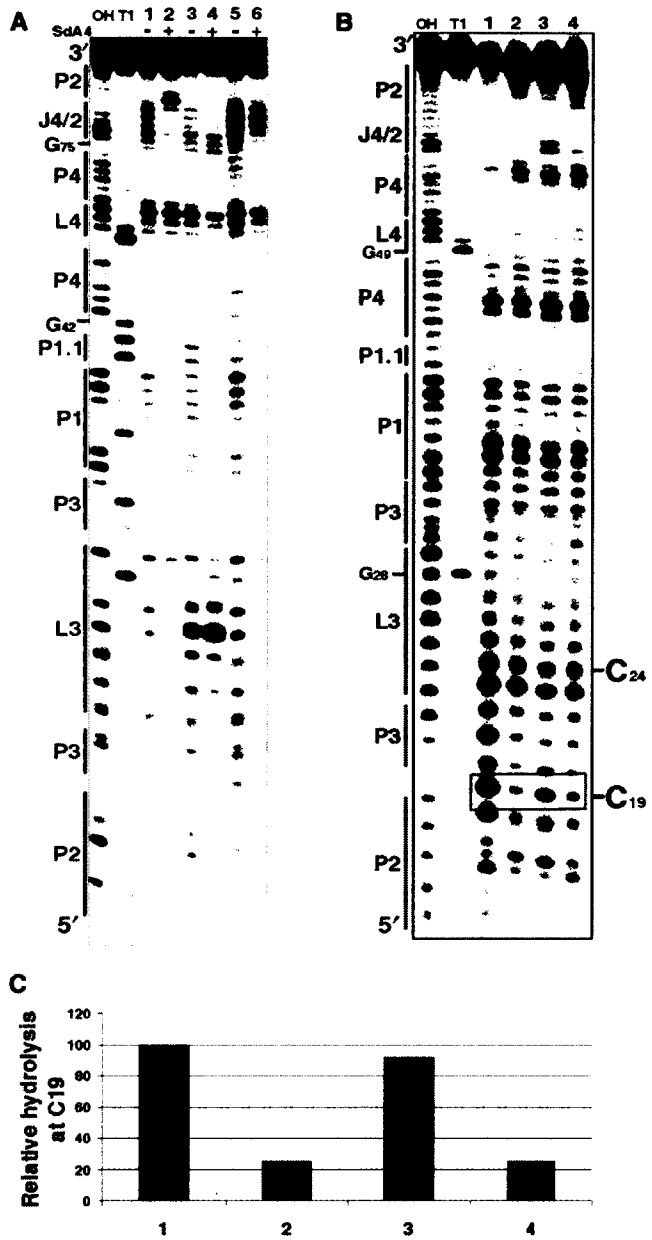


Figure 4. Chemical and enzymatic probing of the bottom of the P2 stem. (**A**) Autoradiogram of a 10% PAGE gel of in-line probing performed on 5′-end-labelled wild-type and mutated *trans*-acting ribozymes. Both alkaline and RNase T1 hydrolyses of the wild-type ribozyme were performed in order to determine the location of each position (lanes OH and T1, respectively). In-line probing of the wild-type ribozyme (1, 2), the $RzC_{24}U$,$C_{25}U$,$G_{40}U$,$G_{41}U$ (3, 4) and the $RzA_{78}U$,$A_{79}U$ (5, 6) mutants are shown. The experiments were performed either in the absence (−) or the presence (+) of the SdA4 analogue. The secondary structure motifs are identified on the left. (**B**) Autoradiogram of a 10% PAGE gel of RNase V1 probing performed on 5′-end-labelled wild-type and mutated *cis*-acting ribozymes. Both alkaline and RNase T1 hydrolyses of the wild-type sequence were performed in order to determine the location of each position (lanes OH and T1, respectively). Lanes 1 to 4 correspond to the wild-type sequence ($C_{19}G_{81}G_{80}$) and the mutants RzC_{19}.$G_{81}A$.$G_{80}A$, RzC_{19}.$G_{81}AG_{80}$ and $RzC_{19}G_{81}G_{80}A$, respectively. The positions of the C_{19} and C_{24} (used to establish the relative level of hydrolysis) are indicated on the right. (**C**) Histogram of the relative levels of RNase V1 hydrolysis of C_{19} for each ribozyme.

be less susceptible to hydrolysis, supporting the stacking of these stems into one helix (3). Moreover, the nucleotides of both the L3 loop and J4/2 junction were less hydrolysed, indicating the rearrangement of these single-stranded domains into a more compact catalytic core. However, the most obvious difference in the presence of the substrate was observed at the bottom of the P2 stem: the residue G_{81} became highly susceptible to the in-line attack, indicating that if it is indeed base paired in the absence of the substrate, it switches into a single-stranded conformation in its presence. Similar observations were made when the in-line probing was performed in the presence of the 3′ cleavage product that bound to the P1 region (data not shown). Moreover, terbium-mediated footprinting results also support the occurrence of a rearrangement of the bottom of the P2 stem (21).

It is well established that substrate binding triggers many folding rearrangements, the most important of which is the formation of the P1.1 stem between the L3 loop and the J1/4 junction (3,19). In order to verify if the formation of the C_{19}–G_{80} bp occurred either immediately after the annealing step, or later in the folding pathway, in-line probing of a mutant ribozyme known to prevent the formation of the P1.1 pseudoknot was performed. In this mutant the two G–C bp forming the P1.1 pseudoknot (i.e. G_{40}–C_{25} and G_{41}–C_{24}) are replaced by four uridines (i.e. U_{24}, U_{25}, U_{40} and U_{41}). This mutant binds the substrate, but is completely deprived of cleavage activity (7). The banding pattern obtained upon in-line probing was virtually identical, regardless of the absence or the presence of the SdA4 analogue, with the exception of within the substrate binding region (Figure 4A, lanes 3 and 4, respectively; see also Supplementary Figure 1B). With this mutant, the presence of the SdA4 did not trigger a significant folding rearrangement, and G_{81} remained insensitive to the in-line attack. This result indicates that G_{81} is most likely base paired to C_{19} prior the formation of the P1.1 pseudoknot, and that only subsequently does it become single-stranded.

Next, in order to gain more confidence for the conformational rearrangement of the bottom of the P2 stem, enzymatic probing was performed using RNase V1, an enzyme that specifically cleaves double-stranded nucleotides. This experiment was performed using a 5′-end-labelled 3′-product derived from the *cis*-acting self-cleaving strand of either the wild type or the mutated sequences (Figure 4B and C). The banding pattern produced by the wild-type ribozyme is in agreement with the proposed secondary structure (Figure 4B, lane 1). C_{19} was hydrolysed, indicating that it is indeed double-stranded. The three mutants tested exhibited essentially identical banding patterns, with only minor differences being observed. For example, the band pattern of the mutant RzC_{19},$G_{81}AG_{80}$ included additional bands corresponding to positions 76 and 77. These two bands might be resulting from an alternative conformation adopted by the J4/2 region or the J4/2 stacked slightly more on the P4 stem, enough for RNase V1 to be active. The most important differences were observed for C_{19} (Figure 4B and C, lanes 2–4). For the C_{19},$G_{81}A$,$G_{80}A$ mutant, which is unable to form any base pair involving C_{19}. RNase V1 poorly hydrolysed this residue. Conversely, it efficiently hydrolysed the C_{19} of the C_{19},$G_{81}A$,G_{80} mutant, suggesting the formation of the C_{19}–G_{80} bp (Figure 4B and C, lane 3). Finally, the RNase V1 did not hydrolyse the C_{19} of the C_{19},G_{81},$G_{80}A$ mutant (Figure 4B, lane 4). The fact that C_{19} remains primarily single-stranded in this mutant, even though base pairing with G_{81} is possible, is an indication that, upon substrate binding, and especially after the formation of the P1.1 stem, G_{81} is extruded out of the catalytic site even if C_{19} is not allowed to base pair with another nucleotide. Together, these results confirm that C_{19} base-paires with G_{80}, and not with G_{81}.

Redrawing the secondary structure of the antigenomic ribozyme

Prior to this study the J4/2 junctions of both the genomic and antigenomic HDV ribozymes were considered to be different (22). The J4/2 junction of the genomic self-cleaving sequences is composed of four nucleotides, $C_{75}G_{76}A/G_{76}A_{78}$ (Figure 5). This contrasts with that of the antigenomic self-cleaving sequences which is composed of five nucleotides, specifically $C_{76}U/C/A_{77}A/G_{78}A_{79}G_{80}$ (Figure 5). However, when the new C_{19}–G_{80} bp is taken into account, the secondary structure of the antigenomic self-cleaving sequence can be redrawn so as to include a J4/2 junction composed of only four nucleotides (Figure 5C). The similarities are not limited to their length, but also include the nucleotide composition. Among the selected sequences, the base pairing between positions 19 and 80 favours an A_{19}–U_{80} bp (42% AU, 13% UA, 21% GC, 17% CG, 2% GU/UG Wobble bp and 7% mismatches, Table 1), as is observed in all natural genomic variants. Moreover, the identity of the nucleotides located in positions 76–79 in both versions is similar, specifically $C_{76}N_{77}A_{78}A_{79}$ (Figure 5B and C). Although the selection experiment was designed with the antigenomic version of the HDV ribozyme, the selected sequences corresponded to a consensus including both genomic and antigenomic ribozymes.

The switch from $C_{19}G_{81}$ to $C_{19}G_{80}$ occurs late in the folding pathway

The results described above demonstrate that initially C19 is base paired to G_{81} and either simultaneously with, or after the formation of, the P1.1 pseudoknot this base pair is disrupted in order to permit formation of the C_{19}–G_{80} bp (Figure 4A). The formation of the P1.1 pseudoknot has been demonstrated to be the limiting step of the HDV folding pathway, in addition to being critical for the folding of the complete J4/2 junction (3,8). The folding of the J4/2 junction involves the formation of a ribose zipper, a trefoil turn and the positioning of the catalytic cytosine within the catalytic core (C76), three features that received physical support from X-ray diffraction, nuclear magnetic resonance (NMR) and fluorescence spectroscopy studies (3,4,23–25). In order to acquire additional biochemical understanding of the J4/2 junction's folding, we analysed our sequence database

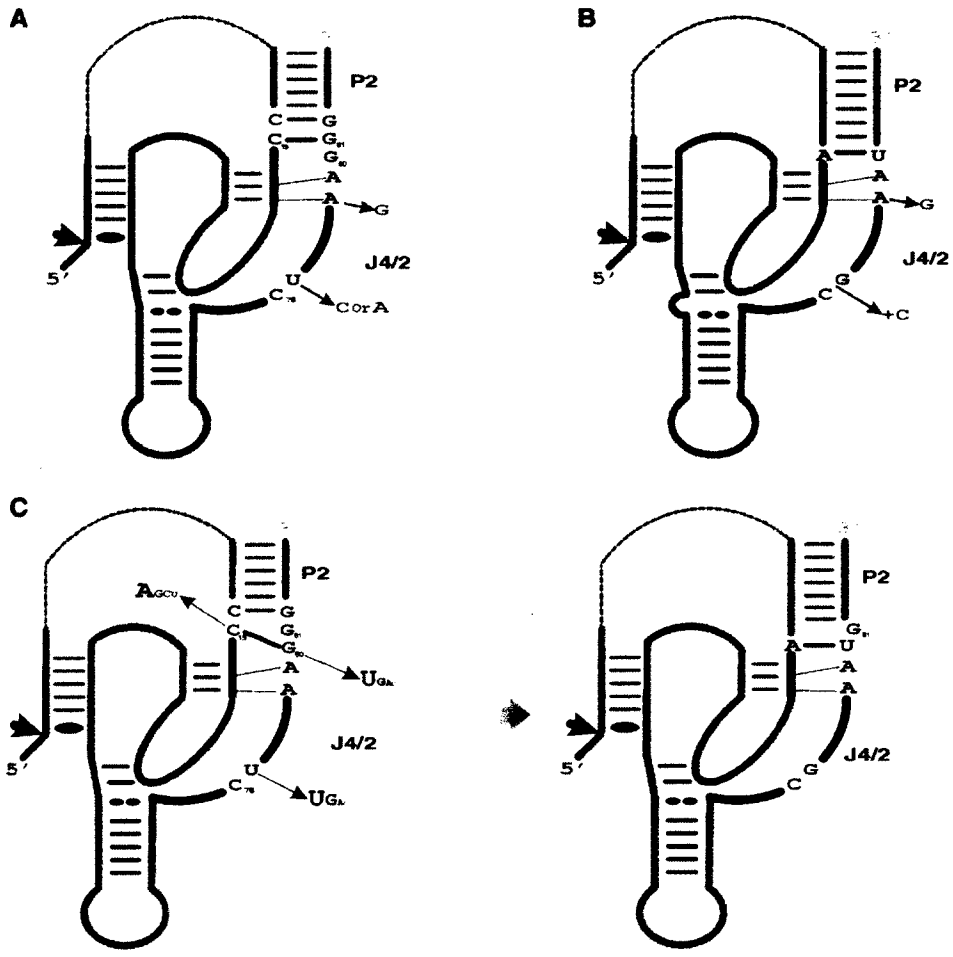


Figure 5. Proposed secondary structures for the HDV self-cleaving RNA motifs. (**A**) and (**B**) are the structures proposed for the antigenomic and genomic motifs prior to the present study. The natural sequence variants reported to date for the J4/2 junction are indicated. (**C**) The novel proposed secondary structures for the antigenomic RNA motif including the C_{19}–G_{80} bp. The differences between both structures are the representations of the C_{19}–G_{80} bp. The sequence variations of the J4/2 junction retrieved via the *in vitro* selection are illustrated. The arrows indicate the cleavage sites.

in terms of these structural features and then synthesized mutated *trans*-acting ribozymes in order to evaluate their cleavage activities (Figure 6).

The ribose zipper is formed by several tertiary interactions between the residues of the G–C bp of the P3 stem and the adenosines, A_{78} and A_{79}, of the J4/2 junction (3,26). According to the selected self-cleaving sequences, A_{79} was perfectly conserved, while A_{78} was highly conserved and could be substituted for by G_{78} in a small number of variants (see Supplementary data Table 2). This mutation has been observed in several natural variants (Figure 5A and B). An important point to bear in mind is that both adenosine and guanosine moieties possess a nitrogen group in position 3 which is involved in the interactions that form the ribose zipper (3). The substitution of both adenosines by two uridines completely abolishes the cleavage activity of the *trans*-acting ribozyme (Figure 6, mutant Rz-A_{78}U,A_{79}U), confirming the importance of the ribose zipper. In order to learn more, the in-line probing of this mutant was performed (Figure 4A, lanes 7 and 8). Upon the addition of the substrate, the P1.1 pseudoknot is formed as shown by the absence of hydrolysis of the corresponding residues (i.e. C_{24}, C_{25}, G_{40} and G_{41}). Interestingly, G_{81} remained intact while G_{80} was hydrolysed, regardless of the presence or not of the substrate. This indicates that the formation of the C_{19}–G_{80} bp requires the formation of the ribose zipper. In order to investigate whether or not the distance between the C_{19}–G_{80} bp and the ribose zipper is important, a mutant with a uridine inserted between these two structural features was synthesized. The resulting ribozyme exhibited a cleavage activity only three-fold less than that of the wild type, suggesting that the distance is not significant (Figure 6, Rz – U^{+}_{79-80}).

We noted that only 10% of the selected sequences included an A_{80}. The presence of an adenosine in position 80 decreases the cleavage activity of the ribozyme in question (Supplementary data Table 2), an effect that was more dramatic when base pairing was possible between positions 19 and 81. For example, the mutant ribozyme

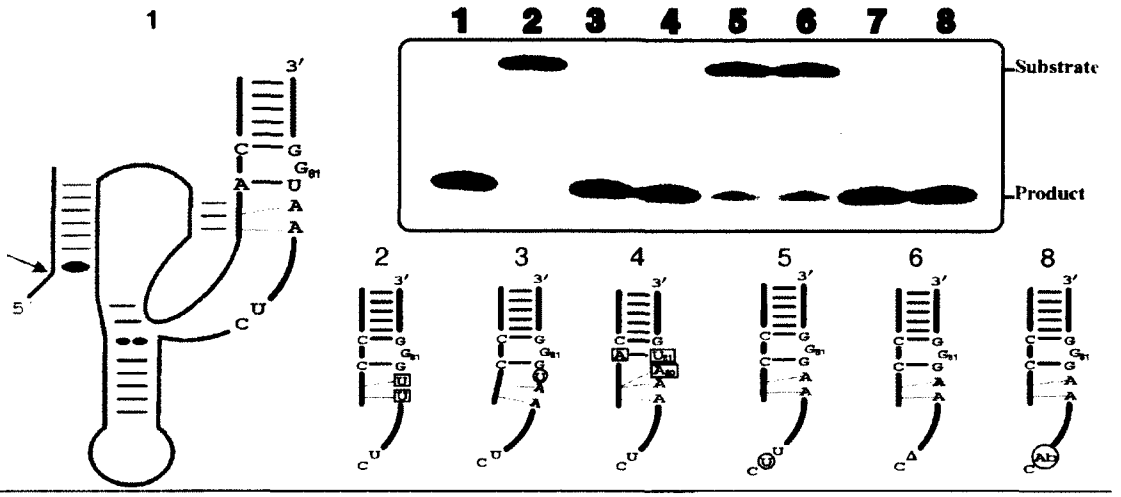


Figure 6. Cleavage activity assays of various mutants of the J4/2 junction. The inset shows a typical autoradiogram of a PAGE gel for a 30 min reaction for each ribozyme tested. The secondary structure of the wild-type ribozyme is illustrated entirely on the left, while only the P2 stem, J4/2 junction and the most relevant nucleotides are illustrated for the mutants. The mutations and insertion are denoted by squared and circled nucleotides, respectively, while deletions are indicated by a triangle (△). The lane numbers at the top of the autoradiogram correspond to the number of each ribozyme: 1, wild-type; 2, $RzA_{78}U,A_{79}U$; 3, $Rz{-}U^{-}{}_{79\ 80}$; 4, $Rz{-}C_{19}A,G_{81}U.G_{80}A$; 5, $Rz{-}U^{-}{}_{76\ 77}$; 6, Rz–△77; 7 is the original bimolecular construct (RzA–RzB); and, 8, the bimolecular construct including an abasic residue (Ab) in position 77 (RzB–Ab77). The positions of the substrate and product are indicated adjacent to the gel. The rate constants (k_{obs}) for each ribozyme are indicated below the gel.

$C_{19}A,G_{81}U.G_{80}A$ exhibited a cleavage activity that was seven-fold less than that of the wild-type ribozyme (Figure 6, wt/$C_{19}A,G_{81}U.G_{80}A$, k_{obs} of 0.85 and 0.12 min^{-1}, respectively). One possible explanation for this effect is that the presence of a single-stranded A_{80} leads to the formation of a ribose zipper that includes an $A_{80}A_{79}$ pair instead of the $A_{79}A_{78}$ one. If this is indeed the case, then the catalytic cytosine would be displaced and the J4/2 junction would contain an additional nucleotide in the trefoil turn domain. In order to verify this possibility, a mutant with an additional uridine inserted between the ribose zipper and C_{76} was synthesized (Figure 6, $Rz - U^{+}_{76-77}$). This mutant exhibited a cleavage activity that was dramatically reduced, in the order of 100-fold as compared to that of the wild-type ribozyme (0.009 min^{-1} compared to 0.85 min^{-1}). This clearly shows that the distance between the ribose zipper and the catalytic C_{76} is very important. In other words, the trefoil turn cannot be altered so as to accommodate an additional nucleotide. The deletion of the residue (U_{77}) caused an even more dramatic effect (Figure 6, Rz–△U_{77}, k_{obs} of 0.01 min^{-1}), showing that this residue is crucial for the formation of the trefoil turn. The database of selected self-cleaving sequences showed that the identity of the nucleotide at position 77 is not important as all four bases were found at this position (see Supplementary data Table 2). To further support this idea U_{77} was replaced by an abasic nucleotide. To do so we adopted a bimolecular ribozyme designed by opening the wild-type ribozyme within the L4 loop, thus generating two pieces: RzA that includes the nucleotides 14–48; and, RzB, which corresponds to nucleotides 67–88 (Figure 1), (27). The ribozymes formed by an RzB strand with either a U_{77} (Figure 6, RzA–RzB) or an abasic residue (RzB–Ab_{77}) exhibited essentially identical cleavage activities (Figure 6; k_{obs} of 0.63 min^{-1} ±0.02 and 0.43 min^{-1} ±0.015 for the RzA–RzB and RzB–Ab77, respectively), confirming that the presence of the base moieties in position 77 is essential, but its identity is not. Only the presence of the phosphodiester backbone at this position is important for the formation of the trefoil turn motif. Altogether, these results show the importance of the base pair switching, and provide additional information on the structural environment required for the formation of the ribose zipper, the trefoil turn and the positioning of the catalytic cytosine within the core centre.

Contribution of the homopurine bp

The homopurine bp is located at the top of the P4 stem and is therefore adjacent to the catalytic cytosine (C_{76}) (see Figure 1, positions 42 and 75). As mentioned previously, this structural feature is perfectly conserved in both the natural and the selected variants. However, it is an $A_{42}A_{75}$ bp in most of the selected variants (72%) and a $G_{42}G_{75}$ bp in the natural variants. A similar observation was detected when analysing the sequence variants from another *in vitro* selection of HDV ribozyme (28). This homopurine bp is important to the ribozyme's activity; its replacement by a Watson–Crick bp has been shown to drastically reduce the cleavage activity (15). However, its contribution to the molecular mechanism of the HDV ribozyme remains elusive.

In order to verify if the adoption of the homopurine bp is influenced by the binding of the substrate to the

ribozyme. RNase T1 probing was performed. RNase T1 hydrolyses all guanosines located in single-stranded regions. In the ribozyme alone, the three consecutive guanosines, G_{40}, G_{41} and G_{42}, appeared as being single-stranded (Figure 7A). Upon addition of the SdA4 analogue these guanosines became inaccessible. The P1.1 pseudoknot, which includes G_{40} and G_{41}, is formed in addition to the homopurine bp (Figure 7A). Thus, the binding of the substrate is required for the formation of these structural features. Several mutants were designed in order to probe the effect of homopurine bp formation on the secondary structure of the ribozyme, especially on the positioning of the catalytic C76. RNase T1 probing of a mutant ribozyme (RzG75C) including a G_{42}–C_{75} Watson–Crick bp instead of a G_{42}–G_{75} homopurine bp was performed (Figure 7A–C). This mutant is completely devoid of any cleavage activity (15). The resulting banding pattern showed several differences (Figure 7A, lanes 3 and 5), the most striking being at the level of the J1/4 junction: the three consecutives guanosines were not hydrolysed regardless of the presence or absence of the SdA4 analogue. This suggests that even in the absence of the substrate analogue, these residues were already engaged in the formation of base pairs. This observation received additional support from in-line probing data showing that the residues of both the J1/4 junction and the bottom of the J4/2 junction (i.e. C75, C76, U77 and A78) were not hydrolysed (data not shown). Together, these results suggest that the residues of these regions form a double-stranded structure that extends the P4 stem (Figure 7C and D). This suggests to us that a potential contribution of the homopurine bp might be to prevent the formation of such a non-productive structure (i.e. alternative inactive folding). The homopurine bp seems to interrupt an elongation of the P4 stem, and contributes to the conservation of the catalytic cytosine as a single-stranded residue. Results from mutagenesis of the homopurine bp have demonstrated that an $A_{42}A_{75}$ homopurine bp is more active than a $G_{42}G_{75}$ bp, as has been observed previously (15). The presence of an AA homopurine bp appears to limit the number of potential alternative structures that can be formed, which may explain the higher occurrence of $A_{42}A_{75}$, rather than $G_{42}G_{75}$, in the selection performed.

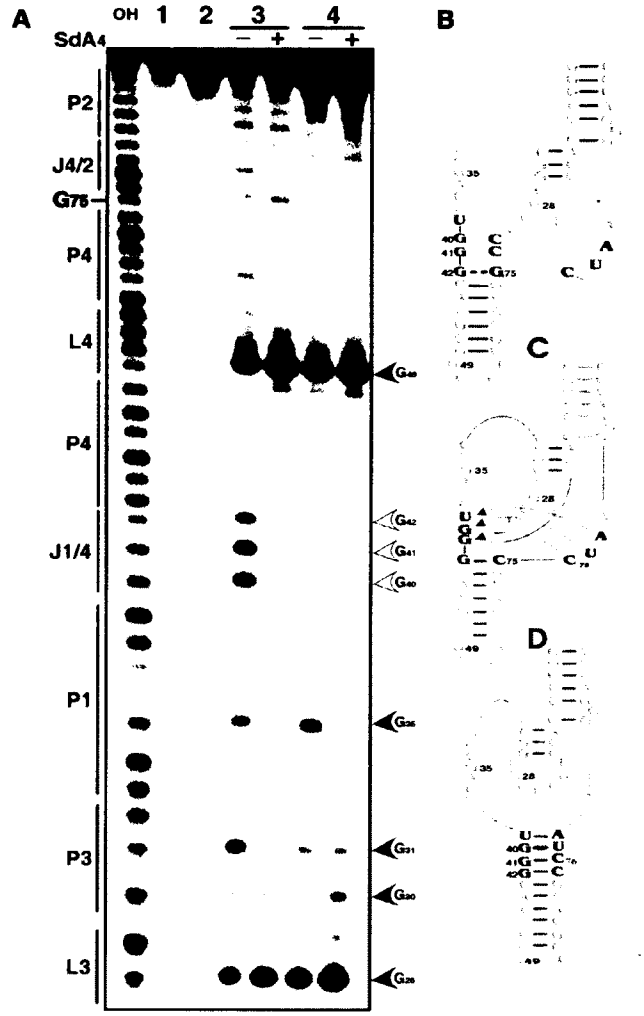


Figure 7. RNase T1 mapping of the wild type and homopurine bp mutant ribozymes. (**A**) Autoradiogram of a 10% PAGE gel of T1 probing performed with 5′-end-labelled wild-type and mutated *trans*-acting ribozymes. Alkaline hydrolysis of the wild-type ribozyme was performed in order to determine the location of each position (lane OH). Lanes 1 and 2 are negative controls (no reaction and no substrate) performed with the wild-type and mutant $RzG_{75}C$ ribozymes, respectively. RNase T1 hydrolysis was performed on both the wild-type ribozyme (lane 3) and the $RzG_{75}C$ (lane 4) either in the absence or the presence of the SdA4 analogue as indicated by the symbols (−) and (+), respectively. The sites of RNase T1 hydrolyses are identified, and intensities of the hydrolyses correlate with the intensities of the arrow heads. The positions of the guanosines are indicated on the left. (**B**) to (**D**) are schematic representation of the nucleotide sequences and secondary structures of the ribozymes. (**B**) is the wild-type ribozyme. (**C**) and (**D**) are two secondary structures for the $RzG_{75}C$ mutant that differs for the J1/4 and J4/2 junctions.

DISCUSSION

The development of an unbiased *in vitro* selection protocol for the isolation of self-cleaving HDV RNA strands, in conjunction with a sequencing effort, permitted the construction of a database containing 119 novel variants. Subsequently, a software programme that evaluates the covariation of nucleotides was developed. The finding that the covariation data supports the base pairs composing both the P1 and P3 stems, as well as that of the homopurine bp, illustrates the usefulness of this software. High CSs were also obtained for some pairs of nucleotides of both the P1 and the P3 stems, suggesting that some interactions must exist between the residues of these two helical regions (e.g. the U_{36} favours the presence of a G_{22}–C_{30} bp in the P3 stem with a CS of +0.47). However, current knowledge does not permit the elucidation of the nature of these putative interactions. NMR studies have already suggested the presence of a magnesium ion located between these two helices (25). Other biochemical experiments will be required in order to establish and characterize these interactions. More importantly, the C_{19}–G_{81} bp at the bottom of the P2 stem was not

supported by the covariation analysis. Conversely, the covariation data indicated the presence of a novel base pair formed between C_{19} and G_{80}. In fact, chemical and enzymatic probing revealed that, upon formation of the P1.1 pseudoknot, the C_{19}–G_{81} bp is disrupted in order to favour the formation of the C_{19}–G_{80} bp. This structural rearrangement occurred as a molecular switch, leaving G_{81} without any contribution to the catalysis.

The P2 stem has been shown to be crucial in the HDV *trans*-acting ribozyme while dispensable for self-cleaving under certain conditions (29, 30). Specifically, an experiment of determination of the 3′ boundary of the self-cleaving antigenomic RNA strand showed that even the ribozyme ending at G_{80} was active (30). However, the level of activity of such a self-catalytic RNA strand remains undefined. It is most likely that the covalent linking between the substrate and catalytic domain is sufficient to permit folding of a significant proportion of the RNA strands into the active conformation. The fact that a *cis*-acting ribozyme ending with the G_{80} at the 3′ end is active supports the importance of the G_{80}–C_{19} base pairing forming the base of P2 stem. According to the older representation of the antigenomic HDV ribozyme, a ribozyme ending with the G_{80} would be deprived of the P2 stem. The identity of the residue at position G_{80} was conserved in all natural variants. Our data showed that the identity of the nucleotide at position 80 affects the level of the cleavage activity; however, the most important feature for this residue is to form a base pair with the residue in position 19; 95% of the self-cleaving sequence included N_{19}–N'_{80} bp (Table 1). It was reported previously that a self-cleaving sequence that did not support the formation of this base pair was not necessarily deprived of all cleavage activity (31). Moreover, we reported recently that the replacement of the G_{80} by a 4-thiouridine residue in a *trans*-acting HDV ribozyme that possess a C_{19} was not detrimental (19). In fact, the cleavage activity was evaluated to be seven-fold smaller (19). In other words, there is some flexibility at position 80. This may be an indication that the switch performed by the G_{80} is probably a consequence of the J4/2 rearrangement instead of being an essential conformation transition. Importantly, the new C_{19}–G_{80} bp sheds light on the contribution of G_{80} that was previously proposed to stabilize the catalytic centre through interaction with the sequence of the P3 stem (30). Rather than being with the P3 stem, this interaction take place with the nucleotide just before the last one of the P2 stem (i.e. C_{19}).

With the formation of the new C_{19}–G_{80} bp, the secondary structure of the antigenomic ribozyme becomes reminiscent of that of the genomic version, that is to say it includes a J4/2 junction composed of only four nucleotides. It has already been suggested that the self-cleaving HDV motifs of both the genomic and the antigenomic polarities adopt a similar global architecture (32), a conclusion supported by our results. The notable differences between the two polarities are primarily limited to the unique single-stranded CAA located within the J1/2 junction of the genomic ribozyme. The reduced number of differences between the two ribozymes strongly suggests that both evolved from a unique ancestral sequence. The various selective pressures exerted along the life cycle of the HDV virus, including its self-cleaving sequences, are most likely responsible for these differences in a manner analogous to that observed between the *in vitro* selected sequences and the natural variants (e.g. AA and GG homopurine bp as well as the C_{19}–G_{80} and A_{19}–U_{80} bp). Interestingly a genome-wide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene (33). The sequence retrieved within the human genome includes neither a G_{80} as found in the genomic version of the self-cleaving HDV strand, nor the single-stranded CAA located within the J1/2 genomic version, this is reminiscent of the antigenomic HDV sequence. Thereby, it looks like a consensus sequence between the genomic and antigenomic sequence variants, supporting the hypothesis that the HDV arose from the human transcriptome (33).

The molecular switch leading to the formation of the C_{19}–G_{80} bp appears to be one feature of a complex structural rearrangement that either occurred simultaneously with, or rapidly following, the formation of the P1.1 pseudoknot. The latter step, which was shown to take place after both the annealing of the substrate to the ribozyme and the docking of the P1 stem within the catalytic centre, has been proposed to be the limiting step of the HDV ribozyme catalysis (8,24). The folding of the P1.1 pseudoknot appears to be the central event that triggers all of the conformational changes leading to the cleavage reaction. This results in the complete folding of the J4/2 junction, including the formation of the C_{19}–G_{80} bp, the ribose zipper (including the $A_{78}A_{79}$) and the trefoil turn ($C_{76}U_{77}$) (Figure 8). Together, the formation of these features leads to the correct positioning of the catalytic cytosine (C_{76}). Furthermore, G_{81} is most likely extruded out of the structure, as is U_{77}. The extrusion of U_{77} is associated with the positioning of C_{76} deeply within the catalytic centre (11). This structural rearrangement might be viewed as two opposing forces causing the exclusion of one nucleotide. In the case of G_{81}, it is tempting to suggest that this is reminiscent of a trefoil turn. Prior to this study, the presence of the structural motifs that form the J4/2 junction, with the exception of the base pair switch C_{19}–G_{80}, were all discovered by high-resolution approaches (3,4,23–25). The present work provides an original biochemical study of these motifs, as well as attempting to determine their order of formation. Moreover, it shows that G_{81} is optional for the ribozyme's catalysis. A mutant lacking G_{81} and including either a C_{19} G_{80} or U_{19}–A_{80} bp exhibited virtually the same cleavage activity as did the wild-type ribozyme. Therefore, the conservation of this guanosine in all natural antigenomic variants is intriguing. Similarly, the GG homopurine bp is perfectly conserved in all of the natural variants, while *in vitro* selection revealed that the most efficient homopurine bp is AA. In addition, we observed several guanosine residues outside of the catalytic centre of the ribozyme (e.g. in the P4 stem-loop) that were conserved in the natural variants, but are not critical to the self-cleavage activity. According to the secondary structure predicted for several HDV RNA genome variants, most, if not all, of these guanosines

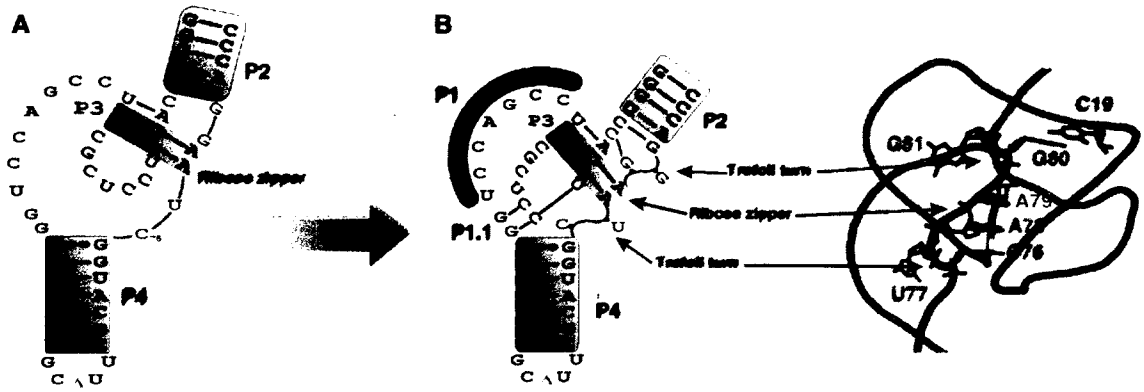


Figure 8. Hypothetical representation of the folding of the J4/2 junction region before and after the formation of the P1.1 pseudoknot. (**A**) Nucleotide sequence and secondary structure of the antigenomic ribozyme characterized in this work. The substrate is represented by the thin line in order to simplify the representation. (**B**) 3D model representation of the antigenomic ribozyme drawn based on the backbone structure obtained from the crystal structure of the genomic version (3). The structural motifs are identified.

appear to form base pairs. Therefore, it is tempting to speculate that a potential contribution of these guanosines would explain their conservation in the natural variants and they favour both, the unfolding of the ribozyme after cleavage and the subsequent formation of the rod-like structure that prevents the self-cleavage of the newly circular HDV RNA strands. Clearly, verifying this hypothesis requires additional experimentation.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Cedric Reymond for his technical assistance. This work was supported by grants from the Canadian Institute of Health Research (CIHR; grant number MOP-44002) to J.P.P. The RNA group is supported by a grant from the Université de Sherbrooke. A.N. was the recipient of a pre-doctoral fellowship from Ministère de l'Enseignement Supérieur et Recherche Scientifique de Tunisie. J.P.P. holds the Canada Research Chair in Genomics and Catalytic RNA. Funding to pay the Open Access publication charges for this article was provided by CIHR.

Conflict of interest statement. None declared.

REFERENCES

1. Been,M.D. (2006) HDV ribozymes. *Curr. Top. Microbiol. Immunol.*, **307**, 47–65.
2. Bergeron,L.J., Ouellet,J. and Perreault,J.P. (2003) Ribozyme-based gene-inactivation systems require a fine comprehension of their substrates specificities: the case of *delta* ribozyme. *Curr. Med. Chem.*, **10**, 2589–2597.
3. FerréD'Amaré,A.R., Zhou,K. and Doudna,J.A. (1998) Crystal structure of a hepatitis delta virus ribozyme. *Nature*, **395**, 567–574.
4. Ke,A., Zhou,K., Ding,F., Cate,J. and Doudna,J.A. (2004) A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature*, **429**, 201–205.
5. Wadkins,T.S., Perrotta,A.T., FerréD'Amaré,A.R., Doudna,J.A. and Been,M.D. (1999) A nested double pseudoknot is required for self-cleavage activity of both the genomic and antigenomic hepatitis *delta* virus ribozymes. *RNA*, **6**, 720–727.
6. Nishikawa,F. and Nishikawa,S. (2000) Requirement of the canonical base pairing in the short pseudoknot structure of genomic hepatitis delta virus ribozyme. *Nucleic Acids Res.*, **28**, 925–931.
7. Deschênes,P., Ouellet,J., Perreault,J. and Perreault,J.P. (2003) Formation of the P1.1 pseudoknot is critical for both the cleavage activity and substrate specificity of an antigenomic *trans*-acting hepatitis *delta* ribozyme. *Nucleic Acids Res.*, **31**, 2087–2096.
8. Ananvoranich,S. and Perreault,J.P. (2000) The kinetic and magnesium requirements for the folding of antigenomic δ ribozymes. *Biochem. Biophys. Res. Comm.*, **270**, 600–607.
9. Bevilacqua,P.C., Brown,T.S., Nakano,S. and Yajima,R. (2004) Catalytic roles for proton transfer and protonation in ribozymes. *Biopolymers*, **73**, 90–109.
10. Das,S.R. and Piccirilli,J.A. (2005) General acid catalysis by the hepatitis delta virus ribozyme. *Nat. Chem. Biol.*, **1**, 45–52.
11. Sefcikova,J., Krasovska,M.V., Spackova,N., Sponer,J. and Walter,N.G. (2007) Impact of an extruded nucleotide on cleavage activity and dynamic catalytic core conformation of the HDV ribozyme. *Biopolymers.*, **85**, 392–406.
12. Nehdi,A. and Perreault,J.P. (2006) Unbiased in vitro selection reveals the unique character of the self-cleaving antigenomic HDV RNA sequence. *Nucleic Acids Res.*, **34**, 584–592.
13. Wu,H.N., Lee,J.Y., Huang,H.W., Huang,Y.S. and Hsueh,T.G. (1993) Mutagenesis analysis of a hepatitis delta virus genomic ribozyme. *Nucleic Acids Res.*, **21**, 4193–4199.
14. Lafontaine,D.A., Ananvoranich,S. and Perreault,J.P. (1999) Presence of a coordinated metal ion in a trans-acting antigenomic delta ribozyme. *Nucleic Acids Res.*, **27**, 3236–3243.
15. Been,M.D. and Perrotta,A.T. (1995) Optimal self-cleavage activity of the hepatitis delta virus RNA is dependent on a homopurine base pair in the ribozyme core. *RNA*, **1**, 1061–1070.
16. Soukup,G.A. and Breaker,R.R. (1999) Relationship between internucleotide linkage geometry and the stability of RNA. *RNA*, **5**, 1308–1325.
17. Nahvi,A., Sudarsan,N., Ebert,M.S., Zou,X., Brown,K.L. and Breaker,R.R. (2002) Genetic control by a metabolite binding mRNA. *Chem. Biol.*, **9**, 1043–1049.
18. Ananvoranich,S. and Perreault,J.P. (1998) Substrate specificity of delta ribozyme cleavage. *J. Biol. Chem.*, **273**, 13182–13188.
19. Reymond,C., Ouellet,J., Bisaillon,M. and Perreault,J.P. (2007) Examination of the folding pathway of the antigenomic hepatitis delta virus ribozyme reveals key interactions of the L3 loop. *RNA*, **13**, 44–54.
20. Ouellet,J. and Perreault,J.P. (2004) Cross-linking experiments reveal the presence of novel structural features between a hepatitis delta virus ribozyme and its substrate. *RNA*, **10**, 1059–1072.

21. Harris,D.A., Tinsley,R.A. and Walter,N.G. (2004) Terbium-mediated footprinting probes a catalytic conformational switch in the antigenomic hepatitis delta virus ribozyme. *J. Mol. Biol.*, **341**, 389–403.
22. Been,M.D. and Wickham,G.S. (1997) Self-cleaving ribozymes of hepatitis delta virus RNA. *Eur. J. Biochem.*, **247**, 741–753.
23. Harris,D.A., Rueda,D. and Walters,N.G. (2002) Local conformational changes in the catalytic core of the trans-acting hepatitis delta virus ribozyme accompany catalysis. *Biochemistry*, **41**, 12051–12061.
24. Pereira,M.J.B., Harris,D.A., Rueda,D. and Walters,N.G. (2002) Reaction pathway of the trans-acting hepatitis delta virus ribozyme: a conformational change accompanies catalysis. *Biochemistry*, **41**, 730–740.
25. Tanaka,Y., Hori,T., Tagaya,M., Sakamoto,T., Kurihara,M. and Uesugi,S. (2002) Imino proton NMR analysis of HDV ribozyme: nested double pseudoknot structure and Mg2+ ion-binding site close to the catalytic core in solution. *Nucleic Acids Res.*, **30**, 766–774.
26. Fiola,K. and Perreault,J.P. (2002) Kinetic and binding analysis of the catalytic involvement of ribose moieties of a trans-acting delta ribozyme. *J. Biol. Chem.*, **277**, 26508–26516.
27. Ananvoranich,S., Fiola,K., Ouellet,J., Deschênes,P. and Perreault,J.P. (2001) Kinetic analysis of bimolecular hepatitis delta virus ribozyme. *Methods Enzymol.*, **341**, 553–566.
28. Legiewicz,M., Wichlacz,A., Brzezicha,B. and Ciesiolka,J. (2006) Antigenomic delta ribozyme variants with mutations in the catalytic core obtained by the in vitro selection method. *Nucleic Acids Res.*, **34**, 1270–1280.
29. Perrotta,A.T. and Been,M.D. (1991) A pseudoknot-like structure required for efficient self-cleavage of hepatitis delta virus RNA. *Nature*, **350**, 434–436.
30. Lee,C.B., Lai,Y.C., Ping,Y.H., Huang,Z.S., Lin,J.Y. and Wu,H.N. (1996) The importance of the helix 2 region for the cis-cleaving and trans-cleaving activities of hepatitis delta virus ribozymes. *Biochemistry*, **35**, 12303–12312.
31. Perrotta,A.T. and Been,M.D. (1996) Core sequences and a cleavage site wobble pair required for HDV antigenomic ribozyme self-cleavage. *Nucleic Acids Res.*, **24**, 1314–1321.
32. Rosenstein,S.R. and Been,M.D. (1996) Hepatitis delta virus ribozymes fold to generate a solvent inaccessible core with essential nucleotides near the cleavage-site phosphate. *Biochemistry*, **35**, 11403–11413.
33. Salehi-Ashtiani,K., Luptak,A., Litovchick,A. and Szostak,J.W. (2006) A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science*, **313**, 1788–1792.
34. Shih,I.H. and Been,M.D. (2002) Catalytic strategies of the hepatitis delta virus ribozymes. *Annu. Rev. Biochem.*, **71**, 887–917.

BJ www.biochemj.org

Biochem. J. (2007) 403, 129–138 (Printed in Great Britain) doi:10.1042/BJ20061216

In vitro selection and characterization of RNA aptamers binding thyroxine hormone

Dominique LÉVESQUE, Jean-Denis BEAUDOIN, Sébastien ROY and Jean-Pierre PERREAULT[1]

RNA Group/Groupe ARN, Département de Biochimie, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec J1H 5N4, Canada

RNA possesses the ability to bind a wide repertoire of small molecules. Some of these binding interactions have been shown to be of primary importance in molecular biology. For example, several classes of mRNA domains, collectively referred to as riboswitches, have been shown to serve as RNA genetic control elements that sense the concentrations of specific metabolites (i.e. acting as direct sensors of chemical compounds). However, to date no RNA species binding a hormone has been reported. Here, we report that the use of an appropriate SELEX (systematic evolution of ligands by exponential enrichment) strategy results in the isolation of thyroxine-specific aptamers. Further biochemical characterization of these aptamers, including mutational studies, the use of transcripts with site-specific modified nucleotides, nuclease and chemical probing, binding-shift assays and CD, demonstrated that these RNA structures included a G-rich motif, reminiscent of a guanine quadruplex structure, adjacent to a helical region. The presence of the thyroxine appeared to be essential for the formation of the structural motif's scaffold. Moreover, the binding is shown to be specific to thyroxine (T4) and tri-iodothyronine (T3), the active forms of the hormone, whereas other inactive derivatives, including thyronine (T0), do not support complex formation. These results suggest that this aptamer specifically binds to the iodine moieties of the thyroxine, a previously unreported ability for an RNA molecule.

Key words: aptamer, riboswitch, RNA ligand, thyroid hormone, thyroxine (T4), tri-iodothyronine (T3).

INTRODUCTION

Several of the original discoveries of biological molecules binding RNA species occurred during the characterization of the self-splicing mechanism of the group I intron (reviewed in [1]). For example, the first step of this mechanism, which is composed of two *trans*-esterification reactions, requires the binding of a GTP that subsequently acts as a nucleophilic group. The amino acid arginine binds to the same site in the intron, leading to inhibition of the splicing reaction [2]. In addition, various antibiotics have been shown to be potential inhibitors of group I intron self-splicing [3]. These initial results revealed the interesting ability of RNA species to bind small molecules [4].

More recently, it has been demonstrated that specific metabolites can directly bind mRNA molecules (reviewed in [5]). Several classes of mRNA domains, collectively referred to as riboswitches and serving as RNA genetic control elements that sense the presence of specific metabolites, act as direct sensors of chemical compounds. Upon interaction with the appropriate small ligand molecule, riboswitch mRNAs undergo a structural reorganization that results in the modulation of the genes that they encode. Riboswitches are known to be responsible for sensing metabolites that are critical for a number of fundamental biochemical processes, metabolites such as coenzyme B_{12}, thiamine pyrophosphate, flavin mononucleotide, *S*-adenosylmethionine, lysine, guanine, adenine and glucosamine-6-phosphate among others [5]. Clearly the ability of RNA molecules to bind a wide repertoire of small ligands is not confined to the test tube, but rather is of primary importance in molecular biology.

In vitro selection, or SELEX (systematic evolution of ligands by exponential enrichment), makes use of a large population of random RNA or DNA sequences as the raw material for the selection of rare functional molecules (reviewed in [6]). This method, basically, consists of sequentially repeating a process that includes the selection of a specific activity (e.g. the binding sequence) coupled to an amplification of either the RNA or DNA molecules possessing this activity. These techniques have broadened our appreciation of the abilities that nucleic acids are capable of. *In vitro* selection has been used to identify aptamers binding a diverse array of small molecules including nucleotides, amino acids, peptides, cofactors, basic antibiotics and transition-state analogues, among others [6,7]. However, RNA species binding either a steroid or a hormone have yet to be reported. We undertook to investigate whether or not T4 (L-thyroxine), as a model hormone, can bind to RNA molecules (Figure 1A). Here we describe the development of a SELEX strategy for the isolation of thyroxine-specific aptamers. The selected aptamers were then characterized further in terms of their binding activities.

EXPERIMENTAL

Preparation of T4-Sepharose

T4-Sepharose was prepared as previously described [8]. Briefly, 6 g of cyanogen bromide-activated Sepharose (Amersham Biosciences) was rinsed with 1 litre of 1 mM HCl and dried. Half of the Sepharose was then mixed with 20 ml of solution containing 50 % (w/v) ethylene glycol, 50 mM $NaHCO_3$ (pH 9.6) and 3 mM of T4 (Sigma). The other half of the Sepharose was prepared in the same solution without T4 (i.e. negative selection). Both solutions were stirred overnight at 4 °C, centrifuged at 5000 *g* for 15 min, washed twice with 50 ml of the solution lacking T4 for 2 h at 4 °C, twice with 50 ml of buffer containing 10 mM Tris/HCl (pH 7.5), 500 mM NaCl and 1 mM $MgCl_2$ for 1 h, and

Abbreviations used: DMS, dimethyl sulfate; G-quartet, guanine quadruplexes; ITP, inosine triphosphate; SELEX, systematic evolution of ligands by exponential enrichment; T0, L-thyronine; T2, 3,5-di-iodo-L-thyronine; T3, 3,3',5-tri-iodo-L-thyronine; T4, L-thyroxine.

[1] To whom correspondence should be addressed (email Jean-Pierre.Perreault@USherbrooke.ca).

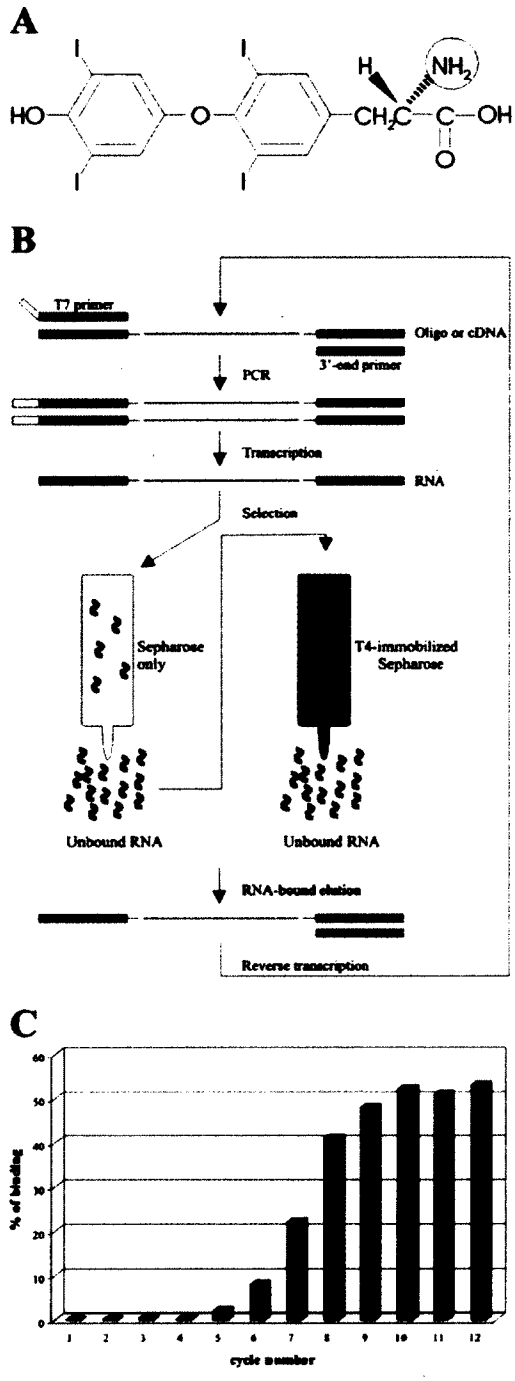


Figure 1 Atomic structure of T4 and selection cycle schematic

(**A**) Atomic structure of T4. The circled amino group is involved in the reaction with the Sepharose. (**B**) Schematic representation of the selection cycle used to isolate T4–aptamer. (**C**) Enrichment profile of the T4–aptamers. The percentage of T4-bound aptamers after each cycle is illustrated in histogram form.

then resuspended in 10 ml of buffer containing 20% (v/v) ethanol, 10 mM Tris/HCl (pH 7.5), 500 mM NaCl and 1 mM $MgCl_2$ and stored at 4 °C. The same procedure was used for the preparation of T3 (3,3',5-triiodo-L-thyronine)-, T2 (3,5-diiodo-L-thyronine)-, 3-iodo-L-tyrosine- and T0 (L-thyronine)-Sepharose (Sigma).

In vitro selection cycle

Synthetic 98-mer oligonucleotides manually synthesized for the random sequence of 58 nt (5'-GAATTCGTCGACGGATCC-N_{58}-CTGCAGGTCGACGCATGCGCCG-3'; Biosource) were amplified by 30 PCR cycles using the T7 primer (5'-TAATACGACTCA-CTATAGGGAATTCGTCGACGGATCC-3') and a 3'-end primer (5'-CGGCGCATGCGTCGACCTGCAG-3'). The PCR products were purified using the QIAquick PCR purification kit (Qiagen).

Purified PCR products were transcribed in 100 μl reactions containing 27 units of RNA guard (Amersham Biosciences), 80 mM Hepes/KOH (pH 7.5), 24 mM $MgCl_2$, 2 mM spermidine, 40 mM dithiothreitol, 4mM of each of ATP, UTP, CTP and GTP, 30 μCi of [α^{32}P]UTP (3000 Ci/mmol; New England Nuclear), 0.01 units of yeast pyrophosphatase (Roche Diagnostic) and 10 μg of purified T7 RNA polymerase at 37 °C for 4 h. The reactions were stopped by the addition of 5 units of RNase free DNase I (Promega) and incubation at 37 °C for 30 min. The resulting products were extracted twice with phenol/chloroform, the nucleic acids precipitated, and subsequently dissolved in 50 μl of water and 25 μl of formamide dye buffer [97.5% (v/v) formamide, 0.025% Xylene Cyanole and 0.025% Bromophenol Blue] prior to being fractionated by 8% (w/v) denaturing (7 M urea) PAGE (acrylamide/bisacrylamide, 19:1, w/v) in buffer containing 45 mM Tris/borate (pH 7.5) and 1 mM EDTA. Products were visualized by UV-shadowing, and the bands corresponding to the full-length RNAs were excised. The transcripts were eluted by a solution of 0.5 M ammonium acetate and 0.1% SDS from the gel slices overnight at 4 °C. The resulting products were passed through Sephadex G-50 spun columns (Amersham Biosciences) and the RNA precipitated. The quantity of RNA was determined by spectrophotometry at 260 nm.

Radiolabelled transcripts (300 pmol) were dissolved in 100 μl of loading buffer [10 mM Tris/HCl (pH 7.5), 500 mM NaCl and 1 mM $MgCl_2$], heated at 65 °C for 2 min and then cooled to room temperature (22 °C). The mixtures were then loaded on to columns containing a 0.2 ml bed column of Sepharose without T4 and incubated for 20 min at room temperature. The columns were rinsed with 300 μl of loading buffer, and the eluates (unbound RNAs) were then loaded on to a second column containing a 0.2 ml bed column of T4-Sepharose and incubated for 20 min at room temperature. The columns were then rinsed with 1.2 ml of loading buffer. Bound transcripts were eluted by either 600 μl of ultrapure water (for cycles 1–6) or 300 μl of 5 M urea (cycles 7–12). Eluted RNAs were ethanol precipitated in the presence of 20 μg of glycogen (Roche Diagnostic). The resulting pellets were conserved for the subsequent reverse transcription reactions. The percentages of RNA bound to the column were calculated as follows: 5 μl of the solutions were recovered both before the loading onto the T4-Sepharose and after the elution, and the radioactivity measured using a scintillation counter.

The pellets from the column eluates were dissolved in 9 μl of ultrapure water and reversed transcribed using Superscript II reverse transcriptase (Invitrogen) as recommended by the manufacturer. The reactions were stopped by the addition of RNase A (10 μg) and incubation at room temperature for 5 min. Aliquots (6 μl) of the resulting products were used as cDNA sources for the PCRs of the subsequent selection cycles.

Cloning and sequencing

PCR products of various cycles were cloned into pGEM-T vector as recommended by the manufacturer (Promega). Several clones were sequenced using the T7 sequencing kit (USB), and the sequences were aligned using the Clustal multiple alignment program [9], followed by minor manual readjustments.

In vitro transcription and ^{32}P labelling of T4-aptamers

A large collection of aptamer variants was synthesized. Briefly, pairs of complementary and overlapping DNA oligonucleotides corresponding to the T7 RNA promoter followed by the full-length aptamer were synthesized and annealed. The second strands were then synthesized by adding 2.5 units of Pwo DNA polymerase (Roche Diagnostic) in a final volume of 100 μl containing 200 μM dNTPs, 10 mM Tris/HCl (pH 8.9), 25 mM KCl, 5 mM $(NH_4)_2SO_4$ and 2 mM $MgSO_4$, followed by five cycles of amplification. The resulting products were purified by extraction with phenol/chloroform twice, then the nucleic acids were precipitated and the DNA templates were *in vitro* transcribed as described above, except that non-radioactive NTP was used. In the case of the aptamers, including either inosine or 7-deaza-GTP, the GTP was replaced by 5 mM GMP and 5 mM of either ITP (inosine triphosphate) or 7-deaza-GTP respectively. After the gel purification and extraction procedures, the RNA aptamers (20 pmol) were dephosphorylated in a final volume of 20 μl containing 200 mM Tris/HCl (pH 8.0), 10 units of RNA Guard and 0.2 units of calf intestinal alkaline phosphatase (Roche Diagnostic) at 37 °C for 30 min. The reactions were purified by extraction with phenol/chloroform twice, and the RNA was ethanol precipitated. Dephosphorylated RNA aptamers (5 pmol) were 5′-end-labelled in a final volume of 10 μl containing 3.2 pmol of [γ-^{32}P]ATP (6000 Ci/mmol), 50 mM Tris/HCl (pH 7.5), 10 mM $MgCl_2$, 50 mM KCl and 3 units of T4 polynucleotide kinase (Amersham Biosciences) at 37 °C for 30 min. The reactions were stopped by the addition of 5 μl of formamide dye buffer, and the mixtures fractionated through denaturing 8 or 12 % (w/v) polyacrylamide gels. The bands containing the appropriate 5′-end-labelled RNAs were excised, and the nucleic acids recovered as described above. For the preparation of 3′-end-labelled aptamers, the latter were incubated in the presence of [^{32}P]Cp (3000 Ci/mmol) and T4 RNA ligase as described by the manufacturer (New England Biolabs), and then purified as described above.

Binding assays

The binding assays of individual aptamers on T4-Sepharose were performed as described above for the selection. Briefly, both radioactive (200 000 c.p.m.) and non-radioactive (300 pmol) aptamers were pooled, and then loaded on to the column. Several washes were performed, and the quantity of unbound aptamers was determined by ^{32}P counting. At least two independent experiments were performed for the determination of the percentages of bound T4-aptamers.

Chemical probing

Either 5′- or 3′-^{32}P-end-labelled (50 000 c.p.m.) ApT4-A' RNA and 30 μM non-radioactive RNA were added to a 9 μl final volume solution containing 1 mM $MgCl_2$, 30 mM sodium cacodylate (pH 7.0) and 150 mM of LiCl, NaCl or KCl. The solution was kept at room temperature for 20 min before the addition of 1 μl of DMS (dimethyl sulfate; diluted 1:8 in 100 % ethanol) and then incubated at room temperature for a further 20 min. The RNA samples were ethanol precipitated, and the pellets washed twice with ethanol in order to remove all traces of DMS. The resulting pellets were dissolved in 20 μl of 500 mM Tris/HCl (pH 7.5). Sodium borohydride (200 mM; 10 μl) was added to the samples, which were then kept on ice for 5 min in the dark. Next, 10 μl of aniline solution (aniline/glacial acetic acid/water, 10:6:93, by vol.) was added to the samples and the tubes incubated at 60 °C for 10 min in the dark. The aptamers were then ethanol precipitated, fractionated on denaturing 10 % (w/v) PAGE and analysed.

Binding shift assay

A mixture of non-radioactive (3 μM) and radioactive (10 000 c.p.m.) aptamers was incubated at room temperature in a final volume of 10 μl containing 10 mM Tris/HCl (pH 7.5) and 1 mM $MgCl_2$ in the absence or the presence of 150 mM NaCl, LiCl or KCl, and with or without 100 μM of T4. After 1 h of incubation, 2 μl of native gel loading buffer was added and the mixtures, which were then analysed on native 15 % (w/v) PAGE gels (acrylamide/bisacrylamide, 29:1, w/v) in buffer containing 45 mM Tris/borate (pH 7.5) and 1 mM EDTA, with or without 150 mM of the salt used for the incubation. When the incubations were performed in the presence of T4, the gel also included 100 μM of T4. The results were visualized with a PhosphorImager (Molecular Dynamics). In order to determine the equilibrium constant (K_d), the experiments were repeated in the presence of various concentrations (1 to 100 μM) of T4 in both the sample and the gel.

CD

CD measurements were performed with a Jasco J-810 spectropolarimeter. The samples were analysed in quartz cells with pathlengths of 1 cm at 22 °C. Far- and near-UV wavelength scans were recorded from 200–250 nm and from 250–340 nm respectively. All experiments were performed using 5 μM ApT4-A' dissolved in 50 mM Tris/HCl (pH 7.5) either in the absence of monovalent salt or in the presence of 50 mM of NaCl or KCl. When required, 100 μM of T4 was added to the samples. The means of at least three wavelength scans are presented. Substraction of the buffer was not required since control experiments in the absence of RNA sample showed negligible curves.

RESULTS

Selection of T4-specific aptamers

Initially an appropriate SELEX strategy was devised in order to isolate T4-specific aptamers (Figure 1). The 98 nt oligodeoxyribonucleotides used as templates for the first PCR amplification included a randomized domain of 58 positions (i.e. an equal likelihood of all four nucleotides at each of the 58 positions). The selection step included two columns. First, a column of Sepharose alone, permitting a negative selection, was used. The transcripts that passed through this column (i.e. the unbound flow-through), which therefore had no affinity for the support material, were selected. Next, a column of T4-Sepharose was used for positive selection from the above initial fraction. The T4 was immobilized through its amine group (Figure 1A). The concentration of T4-Sepharose was determined to be 7 mM by iodine reactivity [10]. The transcripts in the flow-through from this second column (i.e. those positively selected) were discarded, while those bound to the column were eluted and used in the subsequent cycles. Since the transcripts were ^{32}P-radiolabelled, both the retention of the column and its elution were easily monitored. The proportion of transcripts retained on the T4-Sepharose column in each cycle is illustrated in Figure 1(C). This percentage was negligible for the four initial cycles, and then increased significantly at each cycle up to approx. 50 % by cycle nine.

The aptamers resulting from cycle eight, prior to saturation, were cloned and sequenced in order to avoid any bias that might be created by using a limited set of sequences. Out of a total of 60 clones, 53 were found to include a characteristic UGGAGG

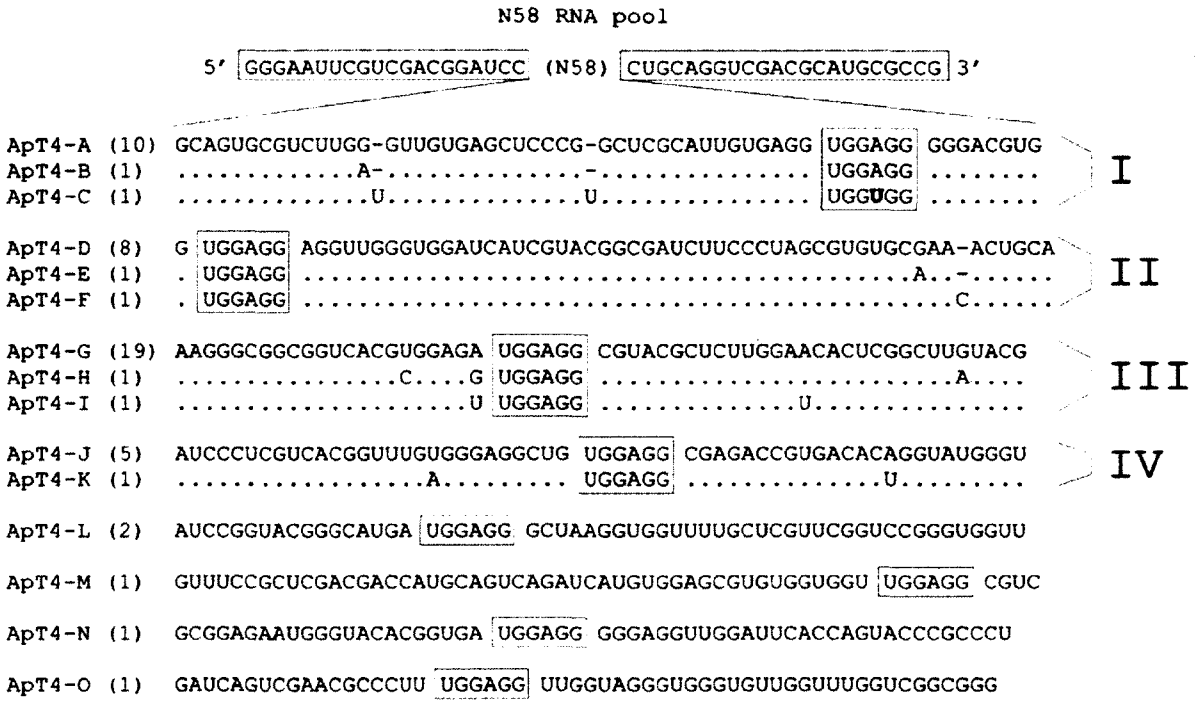


Figure 2 T4–aptamer enrichment

Sequences of the majority of the clones analysed. The conserved UGGAGG sequence is boxed, and the substitution of a U in the ApT4-C is indicated in bold. The frequencies of each clone, after cycle eight, are indicated in parentheses. The groups of sequences (I–IV) are also indicated.

Table 1 Deletion mutants of the ApT4-A aptamer

The sequences of the PCR primers located at both ends of the aptamer, and the conserved UGGAGG sequence, are boxed.

Name	Sequence	Binding %
ApT4-A	5′-GGGAAUUCGUCGACGGAUCCGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUG-CUGCAGGUCGACGCAUGCGCCG-3′	61 ± 5
ApT4-AΔ20nt-5′	5′-GGGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUGCUGCAGGUCGACGCAUGCGCCG-3′	68 ± 4
ApT4-AΔ32nt-5′	5′-GGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUGCUGCAGGUCGACGCAUGCGCCG-3′	17 ± 2
ApT4-AΔ17nt-3′	5′-GGGAAUUCGUCGACGGAUCCGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUGCUGCA-3′	65 ± 5
ApT4-AΔ22nt-3′	5′-GGGAAUUCGUCGACGGAUCCGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUG-3′	31 ± 4
ApT4-AΔ14nt-5′/Δ17nt-3′	5′-GGAUCCGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUGCUGCA-3′	58 ± 3
ApT4-AΔ20nt-5′/Δ22nt-3′	5′-GGGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUG-3′	22 ± 4
ApT4-AΔ20nt-int	5′-GGAUCCGCAGUGCGUCUUGGGUUGUGAG----------GUGGAGGGGGACGUGCUGCA-3′	63 ± 7
ApT4-AΔ30nt-5′int	5′-GGAUCCGCAGUGCGUCUU----------GUGGAGGGGGACGUGCUGCA-3′	22 ± 3
ApT4-AΔ30nt-3′int	5′-GGAUCCGCAGUGCGUCUUGGGUUGUGAG----------ACGUGCUGCA-3′	6 ± 2
ApT4-AΔ-Comp	5′-GGUGCAGCACGUCCCCCUCCACCUCACAAUGCGAGCCGGGAGCUCACAACCCAAGACGCACUGCGGAUCC-3′	< 1 ± 0.5
ApT4-AΔ-db	5′-GGAUCCGCAGUGCGUCUUGGGUUGUGAGCUCCCGGCUCGCAUUGUGAGGUGGAGGGGGACGUGCUGCA-3′ 3′-CCUAGGCGUCACGCAGAACCCAACACUCGAGGGCCGAGCGUAACACUCCACCUCCCCCUGCACGACGUGG 5′	< 1 ± 0.5

sequence box and one possessed an UGGUGG box (Figure 2). Since the conserved UGGAGG box occupies various positions, it cannot be used as a common feature in attempting to produce an accurate alignment. In fact, even considering only the 54 clones possessing either the UGGAGG or the UGGUGG box did not yield a relevant sequence alignment. Moreover, the 54 sequences showed variable abundances. Four groups of closely related sequences accounted for 49 of the 54 sequences. Groups I, II, III and IV include 12, 10, 21 and 6 clones respectively. Ten aptamers from cycle ten, which corresponds to that after the saturation of enrichment, were also sequenced. Only representatives from groups I and II, which always include the UGGAGG box, were retrieved.

Smaller ApT4-A and structural characterization

In order to identify a motif binding T4, we chose to study ApT4-A (Figure 2) for two reasons: it was abundant, and one variation of its UGGAGG box was observed (i.e. UGGUGG). Several deletion mutants were synthesized, and their binding on the T4-Sepharose column was assessed (Table 1). All individual mutants exhibited binding to the Sepharose column deprived of T4, although in very small amount (< 5 %). Therefore any binding activities smaller than 5 % were considered negligible. In general, we observed that the estimated percentages were higher than those observed with the selected populations; for example, ApT4-A bound at 61 ± 5 %. Briefly, the sequences that were used as

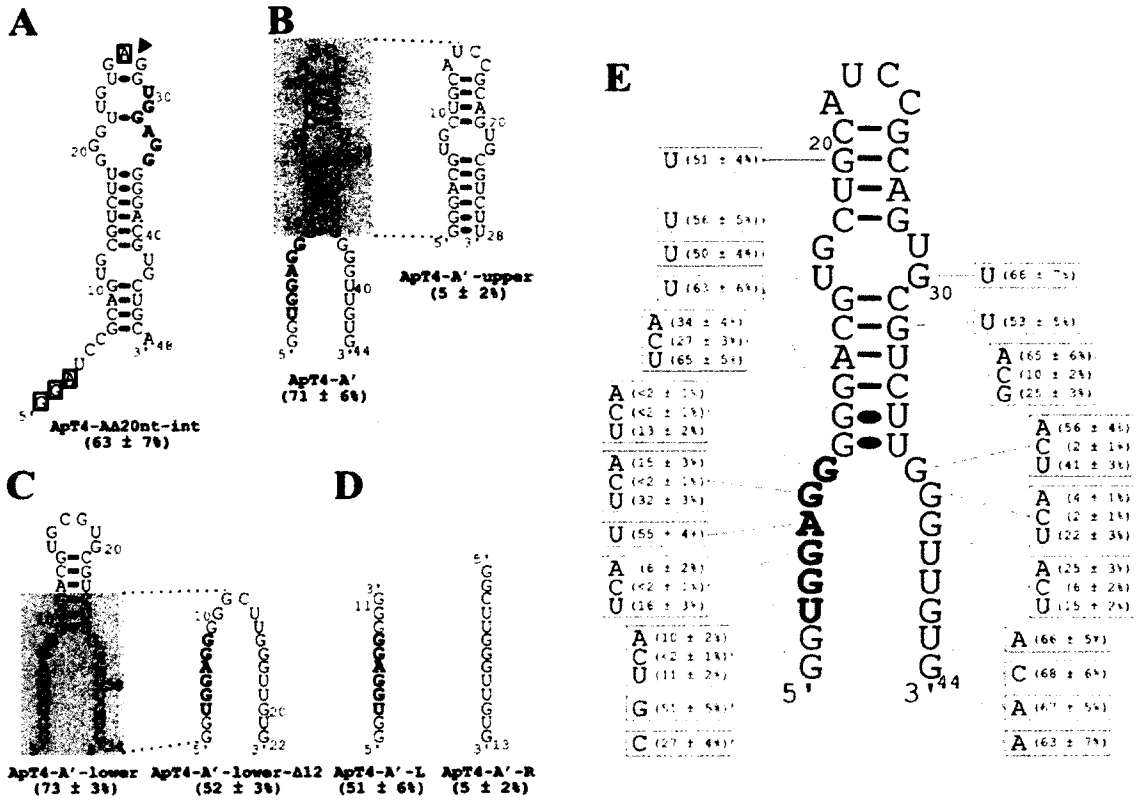


Figure 3 Determination of a minimal structure and important nucleotides for ApT4-A

(**A**) Nucleotide sequence and proposed secondary structure of the deletion mutant ApT4-AΔ20nt-int. The binding activity is indicated in parentheses. The highly conserved sequence UGGAGG is shown in bold. The boxed nucleotides indicate bases that were deleted when designing ApT4-A' [see (**B**)]. The arrow indicates the position where the loop of ApT4-AΔ20nt-int was opened to produce ApT4-A'. (**B**) Nucleotide sequences and proposed secondary structures of ApT4-A' and ApT4-A'-upper. (**C**) Nucleotide sequences and proposed secondary structures of ApT4-A'-lower and ApT4-A'-lower-Δ12. (**D**) Nucleotide sequences and proposed secondary structures of ApT4-A'-L and ApT4-A'-R. (**E**) Nucleotide sequence and proposed secondary structure of ApT4-A'. All point mutations are indicated in the boxes along with the observed binding levels (averages from at least two independent experiments).

the primer-binding sites for PCR can be deleted from the 5′-end without any loss of binding capacity; however, any additional deletion results in a significant loss (i.e. compare ApT4-AΔ20nt-5′ with ApT4-AΔ32nt-5′; Table 1). Conversely, only 17 nt can be removed from the 3′-end without significant loss of binding capacity (compare ApT4-AΔ17nt-3′ with ApT4-AΔ22nt-3′; Table 1). The effects of the various modifications were confirmed with aptamers possessing deletions at both ends (Table 1). Subsequently, in order to further minimize the aptamer, internal deletions were performed (Table 1). When 20 nt from the middle of the aptamer were deleted, the binding ability of the resulting aptamer remained unchanged (ApT4-AΔ20nt-int, $63 \pm 7\%$). However, the further removal of an additional 10 nt significantly reduced the binding percentage (ApT4-AΔ30nt-5′int, $22 \pm 3\%$). Lastly, removal of a further 10 nt, including the UGGAGG box, was noticeably detrimental to binding (ApT4-AΔ30nt-3′int, $6 \pm 2\%$). Thus these results show that it is possible to reduce the aptamer to 48 nt (i.e. ApT4-AΔ20nt-int), and that the presence of the UGGAGG box was essential for efficient binding.

Transcripts complementary to ApT4-AΔ14nt-5′/Δ17nt-3′ were synthesized, and their binding to the column was tested (Table 1). The resulting ApT4-AΔ-Comp did not exhibit any binding activity, indicating that the nature of the sequence is important. Finally, when ApT4-AΔ14nt-5′/Δ17nt-3′ and ApT4-AΔ-Comp were heat-denatured and slowly cooled together in order to favour their annealing into a double-stranded structure, no binding activity was detected. These results indicate that the secondary structure of the aptamer is important for the binding activity.

The most stable secondary structure of ApT4-AΔ20nt-int was predicted using Mfold [11]. Only one putative structure with a ΔG of -18.1 kcal/mol (1 cal $\approx$ 4.184 J) was obtained (Figure 3A). Overall, this structure consists of a hairpin that includes one internal loop and several bulges. In order to verify whether or not the positions of the ends were important for efficient binding, we used a circularly permuted RNA strategy [12]. A phosphodiester bond linked positions 4 and 48, while both the 5′- and 3′-ends were shifted to positions 28 and 26 respectively (i.e. in the upper loop, Figure 3A). During this process the nucleotides G_1–A_3 and A_{27} were deleted, generating the 44 nt ApT4-A' (Figure 3B). The binding capacity of ApT4-A' was found to be similar to that of the previous version (i.e. $71 \pm 6\%$ compared with $63 \pm 7\%$ for ApT4-AΔ20nt-int). The secondary structure of ApT4-A' received physical support from an analysis of RNase H (results not shown). The presence of oligonucleotides complementary to the single-stranded regions of both ends allow for the detection of cleavage product, while oligonucleotide complementary to the sequence of the middle stem did not. Thus ApT4-A' exhibited a single-stranded conserved UGGAGG box adjacent to a double-stranded region.

In order to more precisely define the region responsible for the binding of T4, various mutated aptamers were synthesized using ApT4-A' as a starting point. For example, when the lower portion containing the UGGAGG sequence was removed, the

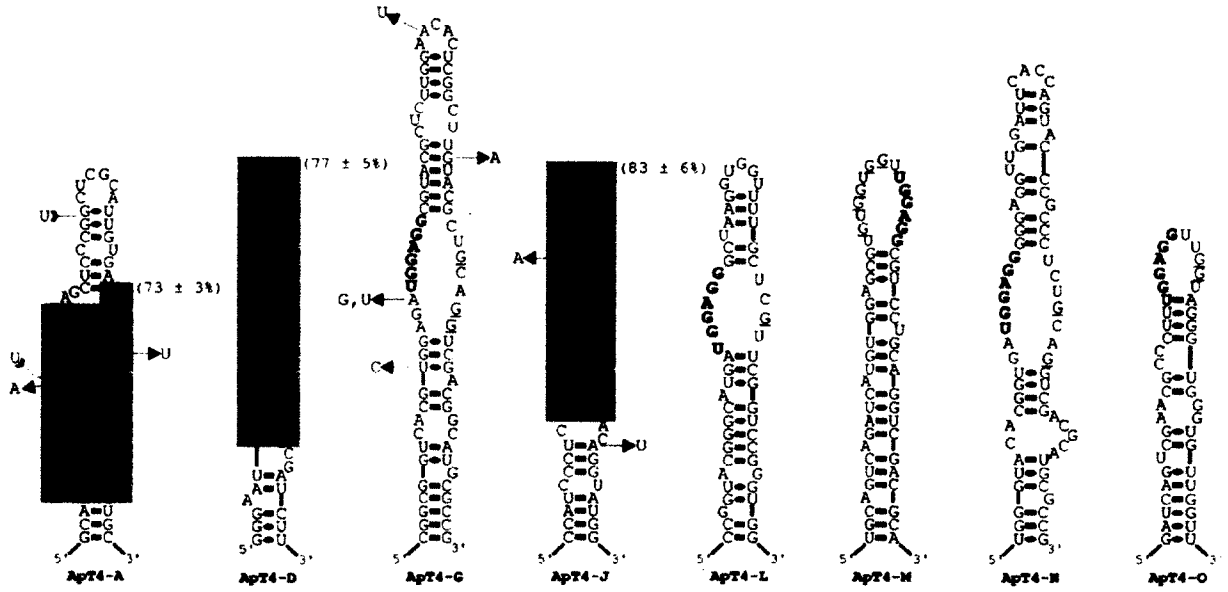


Figure 4 Putative secondary structure of all isolated aptamers

These proposed secondary structures include all requirements known to be important for T4 binding. The lines at both the 5′- and 3′-ends denote sequences that have not been represented in order to simplify the illustration. The natural variants are identified. The UGGAGG sequences are indicated in bold, while the guanosine residues located in front of them are underlined. The grey sections define the smaller versions synthesized and tested for binding and their binding percentages are in parentheses.

residual aptamer (ApT4-A'-upper) was deprived of any binding activity (i.e. $5 \pm 2\%$; Figure 3B). Conversely, when the upper hairpin region was deleted, the resulting aptamer (ApT4-A'-lower) retained full binding activity (i.e. $73 \pm 3\%$; Figure 3C). An additional deletion of 12 nt from the upper domain was also possible, without dramatic loss of binding activity (i.e. ApT4-A'-lower-Δ12, $52 \pm 3\%$; Figure 3C). This surprised us as the resulting aptamer appears likely to be unstructured. Moreover it indicated that the stem, adjacent to the UGGAGG box, had a contribution without being essential (i.e. reduction of 21 %). Other mutated aptamers were produced and showed that the identity of the base pairs composing the stem did not significantly influence the binding activity (results not shown). When the last aptamer was split in two RNA strands, the one including the UGGAGG box exhibited a significant binding activity, while the other strand did not (i.e. ApT4-A'-L and ApT4-A'-R showed binding activity at $51 \pm 6\%$ and $5 \pm 2\%$ respectively; Figure 3D). Together, these results indicate that only the UGGAGG boxed is essential for binding to the T4-Sepharose.

Subsequently point mutations were introduced in many different positions of ApT4-A' in order to identify the nucleotides important for binding (Figure 3E). Only a few mutants resulted in a significant reduction of the binding activity. From the 5′-strand of the aptamer, only mutation of the guanosine residues of the conserved UGGAGG sequence resulted in RNA molecules with low affinities for T4 (i.e. positions 4, 5, 7 and 8). Even the uridine and adenine residues of the UGGAGG can be mutated without significantly reducing the binding percentage (51 % and 55 % respectively, compared with 71 % for the original ApT4-A'). The latter mutation confirmed that the adenosine residue can be replaced by a uridine, as has been observed previously with ApT4-C. Moreover, substitution of the other guanosine residues also results in significant reduction of the binding (e.g. positions 2 and 9). The situation was similar for the point mutations performed on the 3′-strand: only replacement of the three consecutive guanosine residues led to aptamers with low T4 binding (positions 37, 38 and 39). If the aptamer ended with these three guanosine residues, the binding to the T4 was efficient, whereas the replacement of these residues by three adenosines was detrimental to the binding activity (results not shown). Together, these results show that the presence of guanosine residues is the basic building block of this RNA motif. Moreover, these guanosine residues appear to be located in single-stranded regions of the RNA that are adjacent to a stem.

Secondary structures of the different aptamers

We next asked whether or not the structural features of ApT4-A' could be found in the other aptamers. All aptamers were folded using Mfold, and several of the most stable structures were analysed. After minimal manual adjustments, such as the removal of a G · U wobble base pair formed by one of the highly conserved guanosine residues, all isolated aptamers had the ability to fold into a structure reminiscent of ApT4-A' (Figure 4). In some aptamers, the UGGAGG sequence was located within an internal loop, whereas in others it was found in an external loop (compare ApT4-A, -G, -L and -N with ApT4-D, -J, -M and -O). The loops were always relatively large, ranging in size from 10 to 17 nt with an average of 14.4 nt. Furthermore, the location of the UGGAGG sequence in either the 5′- or 3′-strand of the aptamer loops appears to be unimportant. In ApT4-A it is located in the 3′-strand of the internal loop, whereas in ApT4-G, -L and -N it is in the 5′-strand. The same observation was made when analysing the location of the UGGAGG sequence in the external loop. In ApT4-D and -O it is located in the 5′ portion of the loop, while it is located in the 3′ portion in both ApT4-J and -M. Importantly, in all cases the UGGAGG sequence is juxtaposed to a double-stranded region that, for the most part, appears to be stable. Finally, it is noteworthy that all aptamers showed a high predominance of guanosine residues in the strands opposite to the UGGAGG boxes. For example, ApT4-A, -D, -J and -M had five guanosine residues in these positions, whereas only one or two would be expected if

no bias existed. Clearly the presence of guanosine residues is the basic building block of this RNA motif.

In order to verify if the predicted structures were logical, smaller versions of the aptamers ApT4-D and -J, which include the UGGAGG in either the 5′ or 3′ region of the external loop, were synthesized and found to efficiently bind to the T4-Sepharose column (Figure 4, grey sections). Moreover a second *in vitro* selection experiment was performed using the smaller version of the ApT4-J aptamer (results not shown). In this experiment the 16 positions of the loop, and the two corresponding to the adjacent base pair, were randomized and the selection performed as described above. Fifty clones were sequenced and a predominance of guanosine residues was found in the sequence (i.e. an average of 9.8 guanosine residues in the 18 positions), and the randomized regions appeared to form a single-stranded structure. A large proportion of the aptamers were observed to contain the UGGAGG box. Even when the aptamer did not contain this box, several GG-dimers or GGG-trimers were detected in the loop. Together these data revealed that the selection process isolated a single motif that can be located in many different RNA species.

Evaluation of the G-quartet (guanine quadruplexes) hypothesis

G-rich nucleic acid sequences are well known to adopt intermolecular or intramolecular quadruplex structures that are stabilized by the presence of G-quartets (Figure 5A; and reviewed in [13,14]). Since the G-rich single-stranded domain of the aptamers is reminiscent of a G-quartet structure, ApT4-A' aptamers with GTP replaced by either 7-deaza-GTP or ITP were synthesized. In the case of the 7-deaza-GTP, this substitution replaces the nitrogen and its lone pair at position 7 by a CH group, while in the inosine version, the modification removes the NH_2 located at position 2 of the purine ring [15]. With the 7-deaza-GTP, position 7 is incapable of serving as a hydrogen bond acceptor, while the inosine version loses the potential to serve as a hydrogen bond donor from the secondary amine. Consequently, both resulting aptamers should lose four of the eight stabilizing hydrogen bonds per quartet and therefore become unstable. The deaza-GTP would retain the capacity to form Watson–Crick base pairs, but not the inosine version. The ability to bind the T4-Sepharose column was drastically reduced from $71 \pm 6\%$ to $26 \pm 5\%$ and $18 \pm 2\%$ for the deaza- and inosine-aptamers respectively. This supports the hypothesis that the aptamers fold into a structure reminiscent of a G-quartet, rather than into a secondary structure motif.

The hypothesis of a G-rich structure received additional physical support from DMS probing using either 5′- or 3′-^{32}P-end-labelled ApT4-A' aptamers (Figures 5B–5D). The DMS treatment modifies the N7 group of all guanosine residues and was performed either in the absence of salt, or in the presence of LiCl, regardless of whether or not the guanosine residues were present in single- or double-stranded structures. LiCl is well known to suppress G-quartet formation [14,16]. The guanosine residues were also modified in the presence of NaCl, which supports G-quartet formation only under conditions different from those used here [14,16]. Conversely, the seven guanosine residues from the 5′-end and the first three residues from the 3′-end were not modified by DMS treatment when the aptamers were incubated in the presence of KCl (Figures 5B and 5C), which is well known to support G-quartet formation [14,16]. These residues are located in a single-stranded region, and, consequently, should be easily modified by the chemical reaction. Thus the G-quartet was formed only in the presence of KCl and absence of T4. When the experiments were repeated in the presence of T4, the results were similar, with the exception that some protection of the guanosine residues was also observed in the presence of

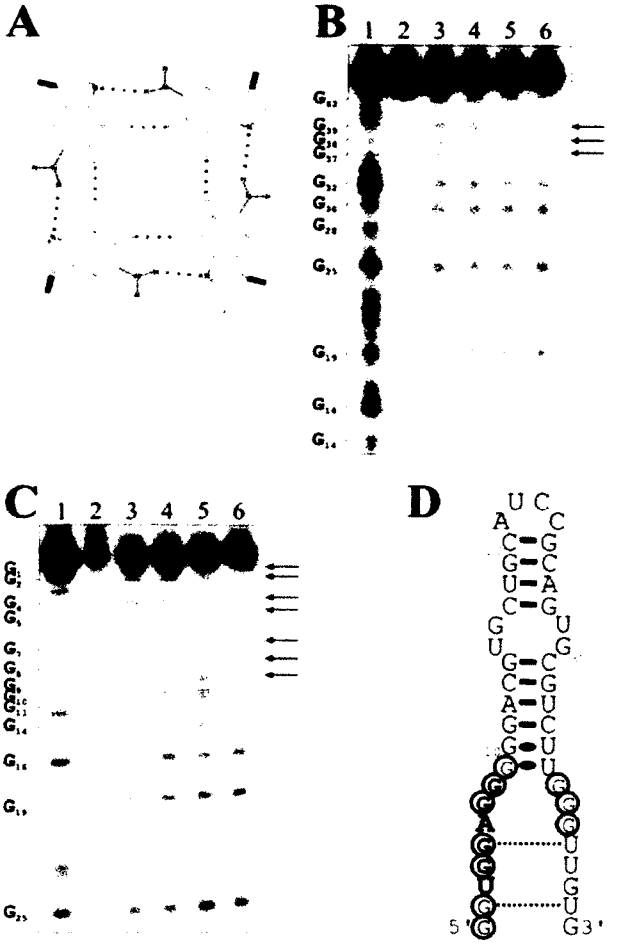


Figure 5 Investigation of a G-quartet-like structure

(**A**) Schematic representation of a G-quartet. The hydrogen bonds between the guanosine residues are indicated by dotted lines. The NH_2 groups linked to the C2 and the N7 positions are circled in grey. (**B**) and (**C**) are autoradiograms of DMS probing of 5′- and 3′-^{32}P-end-labelled ApT4-A' aptamers respectively. Lanes 1: alkaline hydrolyses; lanes 2: the experiments were performed in absence of salt and DMS. Lanes 3–6: show DMS treatments performed either in the absence of salt or in the presence of LiCl, NaCl or KCl respectively. The arrows indicate the resistant guanosine residues. The positions of the guanosine residues are indicated beside the gels. (**D**) Secondary structure and nucleotide sequence of ApT4-A'. The resistant guanosine residues are circled.

NaCl, although at a considerably reduced level compared with that observed in the presence of KCl (results not shown). These observations support the idea that, in the presence of both NaCl and T4, a small proportion of the aptamers fold into a G-rich structure. Several versions of the experiment were repeated in the presence of T4 in order to permit observation of a higher level of protection of the guanosine residues in the presence of the NaCl. Unfortunately these experiments were unsuccessful, most probably due to the limited solubility of the hormone in water (150 μM; [17]), a property that also prevented several other experiments which required higher hormone concentrations. It is also possible that the T4 was modified during the DMS treatment, which would have the effect of limiting its reactivity with the guanosine residues. More importantly, these results are in agreement with the hypothesis of a G-rich structure. Considering the fact that the ApT4-A' aptamers do not include four G-rich segments, quadruplexes should be formed from either two or four RNA molecules.

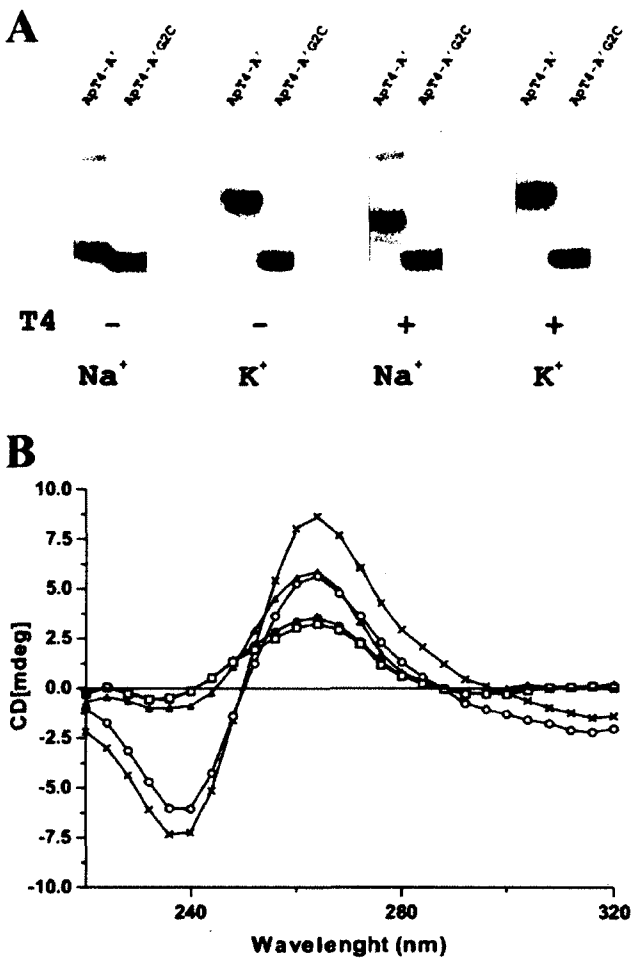


Figure 6 Characterization of the contribution of T4 to the G-quartet-like structure

(**A**) Binding shift assays performed with both the ApT4-A' and ApT4-A'G2C aptamers in the presence of 150 mM of either NaCl or KCl, and with (+) or without (−) 100 μM T4. (**B**) CD spectra performed for the ApT4-A' aptamer under various conditions: no monovalent salt (□), 50 mM NaCl (●), 50 mM KCl (▲) and 100 μM T4 either alone (○) or with 50 mM NaCl (×).

T4 is essential for the formation of the G-rich structure

Subsequently, we asked whether or not T4 was important for the formation of this G-rich structure. In order to answer this question, binding shift assays on native gels were performed. The ApT4-A' aptamer was 5'-end labelled, pre-incubated for 1 h in binding buffer, either with or without T4, and the mixtures then fractionated by native PAGE (15 % gels) either without or with T4 in the gel. In the absence of T4, the transcripts did not shift when the experiments were performed in the presence of 150 mM NaCl, regardless of the aptamer tested (Figure 6A). Similar results were obtained using either a buffer that did not include a monovalent salt, or one containing LiCl, two conditions incompatible with G-quartet formation (results not shown). In fact, in all of these cases, a faint band corresponding to a product with slower electrophoretic migration was observed under all conditions tested. This product most probably corresponds to a misfolded structure adopted by the ApT4-A' and the several mutated versions tested, or to intermolecular products formed by two molecules due to their complementary sequences in the double-stranded region (i.e. as opposed to intramolecular base-pairing). When the experiments were repeated in the presence of both NaCl and 100 μM T4 in the buffer, a shift of the ApT4-A' aptamers was observed (Figure 6A). In fact, almost all of the aptamers showed a slower electrophoretic migration, a property that is characteristic of the formation of an intermolecular G-quartet like structure [18]. Using a range of concentrations in the sample and the gel preparations, the K_d for T4 was estimated to be 50 ± 10 μM. When the experiment was repeated using a mutated version of ApT4-A' in which the guanosine residue at position 2 was substituted by a cytosine residue (ApT4-A'/G2C), no shift was observed (Figure 6A). This confirmed that the slower mobility observed previously was the result of the formation of the G-rich structure. Several other mutated versions were tested, and the results always correlated with the results of binding to the T4-Sepharose: an aptamer that bound the T4-Sepharose shifted on the native gel (results not shown). Finally, the experiment was repeated using the 44nt ApT4-A' and the 22nt ApT4-A'-lower-Δ12nt together. If the formation of the G-like structure is intramolecular, one shifted band for each aptamer should be detected. Experimentally, we observed three predominant shifted bands of different intensities. Using only one radioactive aptamer at a time permitted the detection of intermolecular complexes including only either ApT4-A' or ApT4-A'-lower-Δ12nt, or the two aptamers together, indicating that at least two RNA molecules were involved in the G-quartet-like structure. However, we also consistently detected, in smaller amounts, two other shifted bands that most probably correspond to other complexes formed under these conditions.

When the experiment was repeated in the presence of 150 mM KCl in the buffers, a condition known to favour G-quartet structure, a shift was observed with the ApT4-A' aptamers, regardless of the presence or absence of the T4 (Figure 6A). However, it is important to note that the position of the latter shift was slightly higher than that observed in the presence of NaCl. This might be an indication that different structures were formed depending on whether NaCl or KCl was present. In the presence of KCl, we also accumulated evidence supporting the notion that the G-quartet structures were intermolecular complexes including at least two aptamers (e.g. it was aptamer concentration dependent). More importantly, together, these results suggest that the T4 is essential for the G-quartet-like structure formation when the buffer contains NaCl.

In light of the results described above, we investigated the conformation of the G-quartet structure adopted by the ApT4-A' aptamer using CD. A quadruplex formed by parallel strands is characterized by a long-wavelength positive maximum peak near 265 nm (and a negative peak at 240 nm), whereas a structure including antiparallel DNA strands is associated with a peak near 293 nm [16]. No specific peaks were detected at these wavelengths when the aptamer was incubated either in the absence of monovalent ions or in the presence of NaCl (Figure 6B). Conversely the addition of either KCl or T4 alone in the samples was sufficient to cause detection of a peak at 265 nm, suggesting that these two conditions were sufficient for a proportion of the aptamers to adopt a G-quartet structure. Interestingly, the addition of both the T4 and NaCl to the ApT4-A' yielded a significantly larger peak at 265 nm, indicating that the G-quartet structure is formed by parallel RNA strands. More importantly, it confirmed the essential role of T4 in the formation of the G-quartet-like structure.

Specificity of T4 binding

Initially we investigated whether or not the binding of the aptamer to the column was specific to T4. Columns with immobilized T0, iodotyrosine, T2, T3 and T4 were produced, and the binding of ApT4-A' aptamers to these columns was compared (Figure 7).

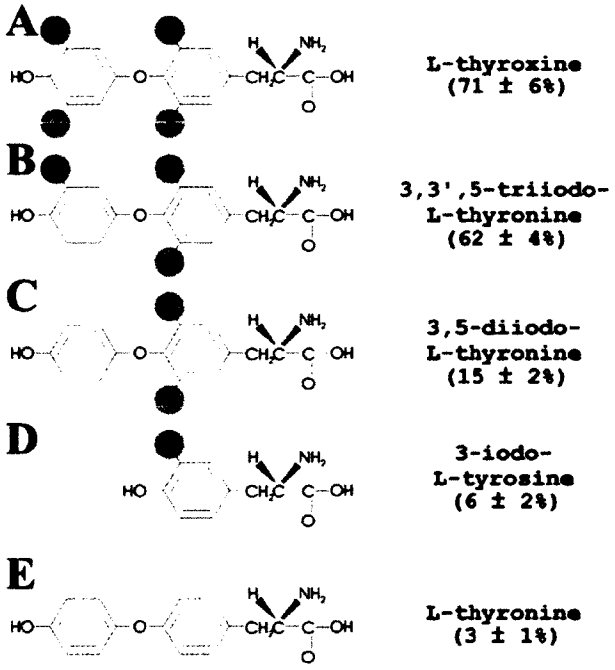


Figure 7 Atomic structure of T4-derivatives and binding activity of the ApT4-A'

(**A**) T4, (**B**) T3, (**C**) T2, (**D**) 3-iodo-L-tyrosine and (**E**) T0.

ApT4-A' bound to the T0, iodotyrosine, T2, T3 and T4 columns at $3 \pm 1\%$, $6 \pm 2\%$, $15 \pm 2\%$, $62 \pm 4\%$ and $71 \pm 6\%$ respectively. In other words, only the T3 and T4 hormones allowed efficient binding, indicating that at least three atoms of iodine must be present for binding to occur. Both the outer and inner rings of T4 appeared to be essential for the binding, and both rings must possess an iodine atom.

DISCUSSION

The designed *in vitro* selection protocol identified several RNA species with the capacity to bind the hormone T4. The mutational analysis and RNase H probing of the ApT4-A' in combination with a comparison of the predicted secondary structures of the various aptamers, led to the identification of several common structural features (Figures 3 and 4). Together, these results show that the presence of guanosine residues is the basic building block of this RNA motif. These guanosine residues (and more specifically the UGGAGG box) were located in single-stranded regions that are adjacent to a double-stranded stem. Mutational analysis demonstrated that RNA species possessing this sequence were able to bind to T4-Sepharose better; therefore it is expected that these aptamers will be more abundant. Similar results were obtained in two other independent selection experiments using different preparations of randomized oligonucleotides (results not shown).

The high guanosine residue content, as well as the fact that the randomized region appears to be unable to fold into a secondary structure, led us to postulate that the RNA species may fold into a G-quartet or a G-quartet-like structure. This hypothesis was confirmed by DMS probing, by T4-Sepharose binding using aptamers synthesized with site-specific modified nucleotides and by electrophoresis binding shift assay experiments (Figures 5 and 6). Together, these experiments provide a strong case in favour of the formation of a G-quartet or G-quartet-like structure that most probably involves either two or four parallel aptamer molecules. Interestingly the ApT4-A' DNA aptamer was also observed to bind to the T4-Sepharose, although at a reduced level (results not shown). Considering that DNA molecules can also fold into G-quartet structures, as observed with the telomeric sequence [19], this provides evidence in favour of such a structure. Some RNA and DNA sequences known to form a G-quartet were tested for their ability to bind to T4-Sepharose; however, at best, only weak binding was observed (e.g. the thrombin-binding DNA aptamer; results not shown). This indicates that G-quartet structures do not have the natural ability to bind T4. One way to reconcile these data is to propose that the sequence of the aptamer is important for the G-quartet formation and that the adjacent base-paired region also contributes to the structure involved in the binding to T4. Preliminary in-line probing experiments support this hypothesis. Specifically, when the probing was performed in the presence of T4 the adjacent helical region appears not to be hydrolysed, whereas in the absence of T4 it is (results not shown). However, the latter observation might also result from the fact that once the G-rich structure is formed between several copies of the aptamer, the stems of each one became parallel and were thus protected from hydrolysis.

The mechanism of the binding of T4 to the RNA aptamer is another intriguing issue. One interesting possibility is the formation of halogen bonds. Halogen bonds in biomolecules can be defined as a short C–X···O–Y interaction in which the X is a carbon-bonded chlorine, bromine or iodine, and the O–Y is a carbonyl, hydroxyl, charged carboxylate or phosphate group [20]. Study of the geometry of halogen bonds in small molecules showed that the interaction is primarily electrostatic, with additional contributions from polarization, dispersion and charge transfer [21]. Although halogen bonds are generally referred to as weak interactions, they have been reported to be exploited in the design of very specific and efficient recognition systems involving proteins [22–24]. For example, the binding between T4 and its transport protein transthyretin has been shown to involve the formation of several I···O halogen bonds [22]. It has been suggested that this large number of short halogen bonds plays an essential role in the recognition of this hormone by its cognate proteins [20]. The demonstration of halogen bonds in nucleic acid structures is limited to only two cases [25,26]. In both of these cases the presence of halogen bonds was shown, by crystallographic study, to take place in complex structures adopted by DNA molecules. More specifically, a Br···O–P halogen bond was shown to be formed in a complex four-stranded junction formed by the oligonucleotide d(CCAGTACBr^{5}UGG) (br^{5}U, 5-bromouridine) [25], whereas an I···O–P short link was detected in a six-stranded complex adopted by the oligonucleotide d(Gi^{5}CGAAAGCT) (i^{5}C, 5-iodocytosine) [26]. To our knowledge such halogens bonds have not yet to be observed in RNA structures either in solution or in crystal form. Therefore if they contribute to the specific binding of T4 to the RNA aptamer characterized here, it would be an original observation. Additional high-resolution structural studies using both NMR and X-ray diffraction should provide definitive proof of G-quartet formation, as well as of the involvement of any halogen bonds between the RNA aptamer and T4.

An interesting question is whether or not aptamers were in fact selected for their ability to bind to T4-Sepharose, or for their ability to fold into a G-quartet-like structure. In this regard, the binding shift assays performed in the presence of NaCl are the most relevant since the initial selection was performed in the

presence of this salt. In the absence of T4, only a negligible fraction of ApT4-A' shifted, whereas in the presence of T4 most of the aptamers formed complexes with a K_d of approx. 50 μM. Since T4 has a limited solubility in water of 150 μM [17], this K_d value is impressive, because it is impossible to fully saturate the complex with T4. Moreover this situation renders the accurate determination of the aptamer–T4 stoichiometry within the complex almost impossible. However, the binding shift experiments suggest that T4 has a role in the formation of the G-quartet-like motif, explaining why all of the isolated aptamers possessed this feature. Moreover, it eliminates the possibility that the G-rich structures were formed in the solution of transcripts before their application on to the column. More likely, the T4 molecules that bound to the column served as scaffolds for the formation of this structure, in a manner reminiscent of the switch role of sodium-potassium in the formation of the DNA G-quartet [19]. Such structural motifs are not only restricted to the telomeric sequence. For example, DNA G-quartets have been proposed to be formed by the human *c-myc* oncogene promoter [27], while RNA equivalents have been found in mRNA and proposed to be important for ribonucleoprotein particle formation and mRNA localization [28,29]. As a result, we were not necessarily surprised to find one more example, although in the present case it has no demonstrated biological relevance.

This work provides an original demonstration that RNA species can specifically bind a hormone. Previously a DNA aptamer has been reported to bind to the T4 hormone [30]; however, the structure of this DNA aptamer bears no relation whatsoever to those of the RNA aptamers isolated here. This difference most probably reflects the different conditions used in both experiments. We do not know whether or not such RNA aptamers occur in natural RNA species found in living cells. If so, it might have a biological importance such as the riboswitch reported to regulate mRNA expression in bacteria [5,31]. Clearly, thyroid hormones are a suitable metabolite with which to search for a potential human riboswitch. Finding equivalent structures within natural mRNAs would most likely lead to a breakthrough in the molecular biology of the thyroid hormones.

We thank Dr M. Bisaillon (Département de Biochimie, Faculté de Médecine, Université de Sherbrooke, Québec, Canada) for access to the CD. This work was supported by grants from the Canadian Institute of Health Research (CIHR) and the Natural Sciences and Engineering Research Council (NSERC) of Canada to J.-P.P. The RNA group is supported by grants from Génome Québec and Université de Sherbrooke. J.-P.P. holds the Canada Research Chair in Genomics and Catalytic RNA.

REFERENCES

1 Cech, T. R. (1990) Self-splicing of group I introns. Annu. Rev. Biochem. **59**, 543–568
2 Yarus, M. (1988) A specific amino acid binding site composed of RNA. Science **240**, 1751–1758
3 Schroeder, R., Waldsich, C. and Wank, H. (2000) Modulation of RNA function by aminoglycoside antibiotics. EMBO J. **19**, 1–9
4 Chow, C. S. and Bogdan, F. M. (1997) A structural basis for RNA–ligand interactions. Chem. Rev. **97**, 1489–1514
5 Tucker, B. J. and Breaker, R. R. (2005) Riboswitches as versatile gene control elements. Curr. Opin. Struct. Biol. **15**, 342–348
6 Wilson, D. S. and Szostak, J. W. (1999) *In vitro* selection of functional nucleic acids. Annu. Rev. Biochem. **68**, 611–647
7 Brody, E. N. and Gold, L. (2000) Aptamers as therapeutic and diagnostic agents. J. Biotechnol. **74**, 5–13
8 Cuatrecasas, P. and Anfinsen, C. B. (1971) Affinity chromatography. Methods Enzymol. **22**, 345–378
9 Thomson, J. D., Higgins, D. G. and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**, 4673–4680
10 O'Kennedy, R., Bator, J. M. and Reading, C. (1989) A microassay for the determination of iodide and its application to the measurement of the iodination of proteins and the catalytic activities of iodo compounds. Anal. Biochem. **179**, 138–144
11 Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. **31**, 3406–3415
12 Nolan, J. M., Burke, D. H. and Pace, N. R. (1993) Circularly permuted tRNAs as specific photoaffinity probes of ribonuclease P RNA structure. Science **261**, 762–765
13 Davis, J. T. (2004) G-quartets 40 years later: from 5'-GMP to molecular biology and supramolecular chemistry. Angew. Chem. Int. Ed. Engl. **43**, 668–698
14 Keniry, M. A. (2001) Quadruplex structures in nucleic acids. Biopolymers **56**, 123–146
15 Kuzmine, I., Gottlieb, P. A. and Martin, C. T. (2001) Structure in nascent RNA leads to termination of slippage transcription by T7 RNA polymerase. Nucleic Acids Res. **29**, 2601–2606
16 Dapic, V., Abdomerovic, V., Marrington, R., Peberdy, J., Rodger, A., Trent, J. O. and Bates, P. (2003) Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. Nucleic Acids Res. **31**, 2097–2107
17 Boulton, D. J., Fawcett, J. P. and Woods, D. J. (1996) Stability of an extemporaneously compounded levothyroxine sodium oral liquid. Am. J. Health Syst. Pharm. **53**, 1157–1161
18 Tang, C. F. and Shafer, R. H. (2006) Engineering the quadruplex fold: nucleoside conformation determines both folding topology and molecularity in guanine quadruplexes. J. Am. Chem. Soc. **128**, 5966–5973
19 Sen, D. and Gilbert, W. (1990) A sodium-potassium switch in the formation of four stranded G4-DNA. Nature **344**, 410–414
20 Auffinger, P., Hays, F. A., Westhof, E. and Ho, P. S. (2004) Halogen bonds in biological molecules. Proc. Natl. Acad. Sci. U.S.A. **101**, 16789–16794
21 Corradi, E., Meille, S. V., Messina, M. T., Metrangolo, P. and Resnati, G. (2000) Halogen bonding versus hydrogen bonding in driving self-assembly processes perfluorocarbon-hydrocarbon self-assembly, part IX. Angew. Chem. Int. Ed. Engl. **15**, 1782–1786
22 Wojtczak, A., Cody, V., Luft, J. R. and Pangborn, W. (2001) Structure of rat transthyretin (rTTR) complex with thyroxine at 2.5 Å resolution: first non-biased insight into thyroxine binding reveals different hormone orientation in two binding sites. Acta Crystallogr. Sect. D **57**, 1061–1070
23 Howard, E. I., Sanishvili, R., Cachau, R. E., Mitschler, A., Chevrier, B., Barth, P., Lamour, V., Van Zandt, M., Sibley, E., Bon, C. et al. (2004) Ultrahigh resolution drug design I: details of interactions in human aldose reductase-inhibitor complex at 0.66 A. Proteins **55**, 792–804
24 De Moliner, E., Brown, N. R. and Johnson, L. N. (2003) Alternative binding modes of an inhibitor to two different kinases. Eur. J. Biochem. **270**, 3174–3181
25 Hays, F. A., Vargason, J. M. and Ho, P. S. (2003) Effect of sequence on the conformation of DNA holiday junctions. Biochemistry **42**, 9586–9597
26 Sunami, T., Kondo, J., Hirao, I., Watanabe, K., Miura, K. I. and Takenaka, A. (2004) Structure of d(GCGAAAGC) (hexagonal form): a base-intercalated duplex as a stable structure. Acta Crystallogr. Sect. D **60**, 90–96
27 Phan, A. T., Kuryavyi, V., Gaw, H. Y. and Patel, D. J. (2005) Small-molecule interaction with a five-guanine-tract G-quadruplex structure from the human MYC promoter. Nat. Chem. Biol. **1**, 167–173
28 Darnell, J. C., Jensen, K. B., Jin, P., Brown, V., Warren, S. T. and Darnell, R. B. (2001) Fragile X mental retardation protein targets G quartet mRNAs important for neuronal function. Cell **107**, 489–499
29 Kostadinov, R., Malhotra, N., Viotti, M., Shine, R., D'Antonio, L. and Bagga, P. (2006) GRSDB: a database of quadruplex forming G-rich sequences in alternatively processed mammalian pre-mRNA sequences. Nucleic Acids Res. **34**, D119–D124
30 Ito, Y., Kawazoe, N. and Imanishi, Y. (2000) *In vitro* selected oligonucleotides as receptors in binding assays. Methods **22**, 107–114
31 Winkler, W. C. (2005) Riboswitches and the role of noncoding RNAs in bacterial metabolic control. Curr. Opin. Chem. Biol. **9**, 594–602

Received 8 August 2006/27 November 2006; accepted 12 December 2006
Published as BJ Immediate Publication 12 December 2006, doi:10.1042/BJ20061216